UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES
ET INFORMATIQUE APPLIQUÉES

PAR
MARIE-CHANTAL DENIS

CONCEPTION ET RÉALISATION D'UN ENTREPÔT DE DONNÉES INSTITUTIONNEL DANS UNE PERSPECTIVE DE SUPPORT À LA PRISE DE DÉCISION

AOÛT 2008

PRÉSENTATION DU JURY

Noms du jury :

M. Sylvain Delisle, directeur de recherche
Département de mathématiques et d'informatique
À l'Université du Québec à Trois-Rivières

M. Mourad Badri, évaluateur
Département de mathématiques et d'informatique
À l'Université du Québec à Trois-Rivières

M. François Meunier, évaluateur
Département de mathématiques et d'informatique
À l'Université du Québec à Trois-Rivières

AVANT-PROPOS

Depuis 1997, l'Université s'est engagée fermement à intégrer les TIC aux diverses réalités du monde dans lesquelles ses étudiants, ses enseignants et ses personnels évoluent. Afin que tous ces membres de la communauté universitaire puissent bénéficier pleinement de toutes les avancés technologiques, l'UQTR a réalisé trois plans directeurs sur l'intégration d'infrastructures informatiques : le plan directeur en enseignement «Les NTIC à l'UQTR» 1997-2000»; le plan d'intégration des TIC dans l'enseignement, l'apprentissage et la gestion académique 2000-2003 et le plan de soutien pédagogique et technologique 2005-2010 – Accompagner les apprentissages et la réussite de l'étudiant.

Résolument branchée, l'Université dispose d'outils technologiques et de communication à la fine pointe des accès intégrés aux logiciels, des développements de contenu de cours et des activités pédagogiques, de la documentation et de l'information de gestion académique. Cette gamme diversifiée d'outils technologiques requiert cependant la nécessité de soutenir leur performance par l'intégration d'équipements et d'infrastructures offrant une exploitation optimale.

Ces installations et outillages seront des précurseurs de leur domaine. Il nous faut en effet concevoir, adapter, planifier, structurer, rédiger, organiser, établir des plans et devis, se ressourcer, procéder, et ce, au niveau de moyens, méthodes, recettes, systèmes, combinaisons, dispositifs, mécanismes, principes et processus. Toutes ces actions et idées constituent les bases mêmes du présent mémoire de recherche.

Bien que nous ne prétendions pas apporter toutes les réponses à toutes les hypothèses de travail, nous allons focaliser notre recherche sur l'élaboration d'une phase initiale et nécessaire à la poursuite du projet et à son implantation à l'UQTR. Préalablement à notre étude, nous avons consulté le «Plan de soutien pédagogique et technologique 2005-2010» rédigé par le «Groupe de travail» chargé de l'élaboration du plan institutionnel de soutien à la pédagogie universitaire et à l'utilisation des TIC en enseignement.

Nous avons aussi examiné la multitude de rapports et de listes demandés en provenance des diverses instances institutionnelles, des requêtes et des besoins exprimés tant par les utilisateurs des systèmes que par les gestionnaires. Cet examen de la situation nous a permis de conclure qu'il était approprié et pertinent que l'Université se dote de nouveaux équipements technologiques répondant aux nouveaux besoins et exigences des utilisateurs de ses systèmes informatisés.

L'observation attentive de notre réalité informatique nous permet de regrouper en quatre (4) catégories nos outils technologiques : outils de soutien à l'enseignement; outils technopédagogiques; outils de gestion académique et outils de communication et de gestion.

Ces équipements, on le sait, deviennent rapidement désuets si les responsables ne suivent pas l'évolution technologique. Dans ce sens, le «Groupe de travail» a procédé à la formulation de recommandations. Notamment, la recommandation #10 mentionne que : *«Que l'Université bonifie la gamme d'outils technologiques mis à la disposition de la communauté universitaire ... »*

Mais, tout équipement matériel, tout système, requière une infrastructure électronique fiable et performante. C'est le cœur de notre préoccupation dans ce mémoire de recherche : présenter une hypothèse de développement informatique répondant aux nouvelles exigences requises exprimées par les utilisateurs qui sont de plus en plus nombreux et de plus en plus rigoureux à l'égard de la performance des systèmes. Donc, notre infrastructure et nos équipements doivent être adaptés à la réalité de notre quotidien et doivent donc être améliorés. Dans ce sens, le «Groupe de travail» considère : «*Que l'Université mette l'accent sur les activités suivantes au cours de la durée du présent plan»* Et plus particulièrement sur : *«Les outils de gestion académiques: développer des outils d'aide à la prise de décision de type tableau de bord pour les gestionnaires académiques».*

Là aussi le «Groupe de travail» considère qu' «*Au fil des années, l'Université a aménagé et a déployé une infrastructure technologique répondant a des objectifs élevés d'intégration des TIC. En outre, grâce a une subvention de la Fondation canadienne pour l'innovation (FCI), l'Université a modernisé son infrastructure réseautique en 2003-2004, afin d'en améliorer le temps de réponse et la performance globale. L'enquête auprès des enseignants et des étudiants révèle un haut taux de satisfaction à l'égard de cette infrastructure et de ces équipements, et le défi est de l'entretenir, de la maintenir à niveau et d'en accroître la pertinence par rapport aux besoins des usagers en enseignement et en gestion académique et administrative. Il faut maintenir de tels services technologiques et faire en sorte qu'ils soient de qualité, performants et conviviaux.*»

Il formule même quelques recommandations notamment sur l'infrastructure et les équipements : «*Que l'Université poursuive le déploiement de ses infrastructures et équipements technologiques et que, dans le but de préserver le haut taux de satisfaction révelé par les sondages auprès des enseignants et des étudiants, elle mise sur l'accessibilité accrue aux équipements, sur leur qualité et leur mise à jour continue, ainsi que sur leur pertinence par rapport aux besoins des usagers, notamment les enseignants et les étudiants*».

Il est donc incontestable que notre projet de recherche s'insère parfaitement dans les créneaux institutionnels. Nous sommes évidemment conscients qu'il s'agit d'un élément de solution parmi la multitude des besoins mais comme le dit l'adage : «Il faut bien commencer au début !»

REMERCIEMENTS

En premier lieu, permettez-moi de remercier monsieur Sylvain Delisle, mon directeur de recherche, pour m'avoir si judicieusement guidée, appuyée et pour cette confiance qu'il m'a accordée. Votre professionnalisme est pour moi une intarissable source d'inspiration.

Je tiens à remercier aussi très sincèrement pour leurs judicieux conseils et pour avoir gracieusement accepté de réviser cet ouvrage, messieurs Mourad Badri et François Meunier, membres du jury.

Je remercie monsieur Michel D. Pépin, mon supérieur immédiat. Avec sa vision à long terme, il a su donner des ailes au projet.

Je remercie les membres de l'équipe de recherche sur le data mining avec qui les discussions furent très constructives. Plus spécialement, monsieur Michel Charest qui est devenu mon collègue au SSPT et bras droit dans ce projet.

Mes hommages les plus respectueux vont également à tout le personnel du Service de soutien pédagogique et technologique de l'UQTR (SSPT). Particulièrement à :

- Madame Naomi Denoncourt, technicienne en informatique, qui a programmé les outils SAT, OAD et SETPS. Elle fut une aide précieuse au projet.
- Messieurs Louis Brouillette et Michel Chênevert, DBA Oracle, qui ont fait un travail remarquable au niveau de l'accès aux «logs» d'Oracle et de l'implantation du CDC (Change Data Capture) d'Oracle.
- Monsieur Georges-Martin Caron pour sa présence d'esprit.
- Messieurs Stéphane Paquet et Jean-Alexandre Beaudet pour l'aide sur l'outil COGNOS 8.
- Monsieur Jean Paquette et madame Liette Pothier, mes fidèles correcteurs.

Merci à tous pour votre support et pour vos encouragements.

Je remercie sincèrement Mme Ofelia Delfina Cervantes Villagomez qui travaille d'une façon remarquable. Elle est une femme d'exception, elle inspire ma façon de travailler.

J'exprime enfin ma gratitude à ma famille pour leur soutien, leur présence et leurs encouragements. À ma mère qui m'a transmis ses valeurs. Elle vit à travers nous.

À mon conjoint qui a fait de ma vie un paradis. Merci de ta compréhension.

Plusieurs autres personnes sont intervenues tout au long de ce projet. Quelques-unes pour de plus courts moments, d'autres plus longuement, mais toutes resteront dans ma mémoire.

Merci à vous !

DÉDICACE

À mes fils
Dany et Charles

Vous êtes ce que j'ai de plus précieux.

Votre maman

CONCEPTION ET RÉALISATION D'UN ENTREPÔT INSTITUTIONNEL DE DONNÉES DANS UNE PERSPECTIVE DE SUPPORT À LA PRISE DE DÉCISION :

Marie-Chantal Denis

RÉSUMÉ

Ce projet vise à concevoir et à appliquer une approche méthodologique afin de construire un prototype d'entrepôt de données de l'UQTR (Université du Québec à Trois-Rivières) et ainsi faciliter la prise de décision par l'élaboration de tableaux de bord présentés à l'intention des gestionnaires. Le projet, constitué d'une partie théorique et d'une seconde partie hautement pratique, est appliqué à un cas de gestion universitaire d'analyse des clientèles étudiantes.

Préalablement à la phase d'analyse et d'historisation, nous avons procédé à la réalisation de l'étude des méthodologies de conception. La méthodologie développée dans ce projet est fondée sur celle utilisée par Ralph Kimball qui préconise une conception de bout en bout sur un seul processus d'affaires «orienté-sujet». La méthodologie ainsi conçue permettra de poursuivre le projet et d'amalgamer les systèmes actuels de l'UQTR successivement afin d'offrir une homogénéité qui permettra aux demandeurs d'obtenir la compatibilité nécessaire à une consultation efficace et simple de l'ensemble des différents types de données. Dans un deuxième temps, l'analyse d'outils existants a été complétée. Cette étude met en relief les caractéristiques des différents outils menant à une première ébauche pour une recommandation d'achat d'une solution informatique pour l'UQTR. Deux types d'outils étaient nécessaires : un outil de chargement de l'entrepôt (ETL : extract-transform-load) et un outil de présentation en ligne des données. Finalement, une proposition d'historisation des définitions de tables a été développée permettant ainsi d'enrichir les métadonnées notamment en ce qui concerne les valeurs manquantes dues à la modification des structures de données dans le temps. Cette étape offre alors l'opportunité d'en expliquer les causes à l'utilisateur.

Une fois les systèmes intégrés à l'entrepôt, les données seront offertes sous différents formats en passant d'un simple fichier à un tableur Excel vers un rapport sur le Web jusqu'à l'OLAP (*On-Line Analytical Processing*). Une fois cette étape complétée, il ne restera qu'un pas à franchir pour permettre l'exploitation de l'entrepôt par le *«data mining»*.

Une des difficultés majeures qu'il nous a fallu résoudre consistait à gérer l'évolution des structures de données des OLTP (On-line transaction processing) et d'en minimiser l'impact sur l'entrepôt de données. L'évolution s'explique par tout ajout, modification ou suppression de tables ou de champs. On a dû prévoir l'explication du contexte au gestionnaire afin de l'informer des variations possibles de ces structures dans le temps qui sont liées aux valeurs manquantes ou inexplicables mais néanmoins présentes dans l'entrepôt et accessibles par celui-ci. Ces valeurs sont des biais qui passeront cependant inaperçues dans les résultats présentés aux gestionnaires à l'intérieur de tableaux de bord.

Une autre difficulté a été d'expliquer aux gestionnaires les effets temporels qui ont influencé la valeur des données ou la disponibilité d'un champ. Par exemple, la grève des chargés de cours qui a fait augmenter le taux d'échec lors de cette session. Privé de cette information, l'utilisateur demeure sans explication quant à un fait empirique mesurable qui influence négativement la moyenne et l'écart-type. Le SystèmE de Perception Temporel des Structures (SEPTS) fut créé à cet effet.

Dans la phase majeure de la «préparation des données», un important problème d'intégrité a été observé. Les anciens systèmes n'utilisaient pas de clé étrangère, ce qui compromettait l'intégrité des données lors de la conversion du schéma ER (entité-relation) en schéma dimensionnel. La conception de deux outils a été réalisée pour gérer l'intégrité et l'extensibilité des définitions des tables : le Système d'Analyse de Tables (SAT) et l'Outil d'Analyse des DDL (OAD).

Par ce projet, nous croyons que la prise de décision sera alors aisément facilitée quoique nous soyons conscients que cette évolution entraînera une nouvelle problématique. Nous aurons ainsi un nouveau défi à relever. Ce défi consistera à réaliser une gestion intelligente de cette masse de données en établissant une approche d'archivage et de réintégration des données à l'entrepôt afin de satisfaire tous les types d'analyses longitudinales ou tout simplement de réduire le volume exponentiel des données de l'entrepôt.

DESIGN AND IMPLEMENTATION OF AN INSTITUTIONAL DATA WAREHOUSE FROM A DECISION SUPPORT PERSPECTIVE

Marie-Chantal Denis

ABSTRACT

The following project has consisted in the design and implementation of a methodology for the realization of a data warehouse targeted for a university business environment. The principle objective of such an endeavor has been to empower decision makers via the use of dashboard technology. More specifically this project, based on both a theoretical and a particularly strong practical background, has been specifically within the context of better managing university student clientele.

The early phase of our project consisted of evaluating various data warehousing design methodologies. In fact, the methodology we have developed is based on Ralph Kimbal's methodology which is based on the principle of realizing dimensional models which are end-to-end business process and subject-oriented. Hence, the design methodology we shall propose herein aims at encouraging scalability as the project continues to evolve and new heterogeneous data sources become available to end-users. Subsequently, a detailed analysis of existing commercial and open-source data warehouse and business intelligence products was carried out with the aim of eventually recommending a suitable solution at UQTR for meeting the business requirements. More specifically, two types of tools were deemed necessary for the project: a back-end extraction, transformation and loading tool (ETC) and a front-end business intelligence framework for presenting data and information to end-users via the Web. Finally, we have proposed a framework for managing table definition and structure changes which ultimately aims at empowering the user with richer meta-data. This can be particularly helpful for end-users by providing helpful hints on missing values which inherently result from data structure changes.

Once the systems were integrated within our data warehouse environment, the data was made available to users either in the form of reports using various data formats (i.e. web report, MS-Excel, etc.) or OLAP analysis. Eventually, once the data warehouse shall have sufficiently evolved, future plans are underway to make data mining technology available to end-users.

One of the key challenges that has arisen during our research is how one should manage the evolution and changes to the table structures within our operational systems (OLTP), in order to minimize the impact on the data warehouse. By « evolution » we imply any addition, modification or removal of a table field. We have had to consider how to present to the end-user all these possible table structure changes and how they relate to the presence of missing values within the data warehouse. The objective has been for such biases or anomalies to be properly managed and have a transparent impact on the business end-user. Another difficulty to manage has been how to inform the end-user of events which could have occurred that can have a direct impact on the business value of the data available for a given table field. For instance, the occurrence of a strike by professors at a university could superficially increase the student failure rate. Without such key information, a business analyst would have no means for properly correcting such a situation and correctly interpreting the associated data with such an event (i.e. biased average and standard deviation measures).

During the most demanding phase of the data preparation stage, a major integration problem was observed. The legacy systems did not make use of foreign key constraints and this had a direct impact on the integrity of the data that was migrated to the data warehouse (i.e. relation model to dimensional model transformation). As such, two tools have been implemented in order to manage the integrity and extensibility of table definitions: a data analysis tool and a data definition language (DDL) analysis tool.

Though the implementation of this project shall be simplified by the use of such tools, new challenges shall arise. For instance, it shall become important to consider the « intelligent » management of large masses of data by making use of archiving and re-integration techniques in order to potentially satisfy many business analysis requirements from end-users.

TABLES DES MATIÈRES

TABLES DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

Dans le but d'alléger le texte, plusieurs abréviations, sigles et acronymes sont utilisés dans ce document. Leur signification est donnée ci-dessous. S'il existe, l'acronyme français est privilégié dans le texte et vous retrouverez son équivalent anglais dans le tableau ci-contre.

Acronyme	Explication	*Traduction*
3NF	Règles de transformation appliquées aux données afin de les convertir en schéma relationnel. Il existe de la 1er à la 4ième forme normale.	
BI	«Business Intelligence»	*IT*
Comptoir d'information	Regroupement de données ciblé sur un ou plusieurs sujets.	*Data mart*
Cube de données	Cube de présentation d'unités pouvant pivoter sur différentes dimensions.	*Data cube*
Data cube	Cube de présentation d'unités pouvant pivoter sur différentes dimensions.	*Cube de données*
Data mart	Regroupement de données ciblé sur un ou plusieurs sujets.	*Comptoir d'information*
Data mining	Technique de fouille des données pour en extraire de l'information complémentaire et des modèles de connaissances explicatifs ou prédictifs.	*Forage de données*
Data set	Jeux de données statique, extraits de l'entrepôt.	*Jeux de données*
Date warehouse	Structure informatique dans laquelle est centralisée un volume important de données consolidées à partir des différentes sources de renseignements d'une entreprise. L'organisation des données est conçue pour que les personnes intéressées aient accès rapidement et sous forme synthétique à l'information stratégique dont elles ont besoin pour la prise de décision.	*Entrepôt de données*
DDL	Data Definition Language : Langage de définition spécifique pour la gestion des structures de données en SQL pour les bases de données.(Ex. : création de tables, ajout d'un index, ...)	
DDS	Decision support system	*SAD*
DOLAP	Dynamic ou Desktop OLAP	

Acronyme	Explication	*Traduction*
DSI	Sous unité du SSPT : développement des systèmes d'information de l'UQTR	
Entrepôt de données	Structure informatique dans laquelle est centralisée un volume important de données consolidées à partir des différentes sources de renseignements d'une entreprise. L'organisation des données est conçue pour que les personnes intéressées aient accès rapidement et sous forme synthétique à l'information stratégique dont elles ont besoin pour la prise de décision.	*Data warehouse*
ER	Schéma entité-relation	
ETC	«Extraction, Transformation et chargement» Outil informatique destiné à extraire des données de diverses sources, à les transformer et à les charger dans un entrepôt de données.	*ETL*
ETL	«Extraction, Transformation and Loading» Outil informatique destiné à extraire des données de diverses sources, à les transformer et à les charger dans un entrepôt de données.	*ETC*
Forage de données	Technique de fouille des données pour en extraire de l'information complémentaire et des modèles de connaissances explicatifs ou prédictifs.	*Data mining*
HOLAP	Technique hybride entre le MOLAP et le ROLAP.	
MOLAP	Technique OLAP optimisée pour l'analyse multidimensionnelle.	
OAD	Outil d'analyse de données dans les tables (Logiciel UQTR)	
OLAP	«On-Line Analytical Processing» Concept de l'informatique décisionnelle, à mi-chemin entre le système d'information et les utilisateurs, permettant de faire des simulations.	
OLTP	«On-line transaction Processing» Systèmes de traitement de transactions en ligne.	
Oracle	Nom d'une base de données très répandue et disponible sur plusieurs plates-formes.	
OSS	Operator support system	
PowerHouse	Cognos PowerHouse est une solution de développement d'applications utilisée par les organisations pour créer des applications d'entreprise de type terminal.	
ROLAP	Technique de modélisation et de stockage des données basée sur une structure relationnelle.	
SAD	Système d'aide à la décision	*DDS*
SADDL	Système d'analyse des DDL (Logiciel UQTR)	

Acronyme	Explication	*Traduction*
Schéma dimensionnel	Schéma représenté par des dimensions et des faits conformes. Ce schéma n'est pas en 3NF, on le dit « dénormalisé »	*Dimensionnal modeling*
Schéma en constellation	Regroupement de schémas en étoile.	
Schéma étoile	Schéma dimensionnel pouvant avoir une forme en étoile avec seulement une table de fait et plusieurs dimensions unitaires.	*Star schema*
Schéma relationnel (schéma ER)	Schéma représenté par des entités et des relations (ER). Ce schéma est normalement en 3NF (3e forme normale).	*Relational schema*
Schémas en flocon	Schéma dimensionnel pouvant avoir une forme en flocon de neige avec seulement une table de fait et plusieurs dimensions pouvant avoir une sous- dimension en relation.	*Snowflake schema*
SEPTS	SystèmE de Perception Temporel des Structures (Logiciel UQTR)	
SGBD	Système de gestion de base de données	
SOLAP	Spatial OLAP	
SSPT	Service du soutien pédagogique et technologique de l'UQTR	
UQTR	Université du Québec à Trois-Rivières	

Utilisation de l'anglais

Dans la mesure du possible, l'expression française sera privilégiée, par contre, puisque plusieurs termes anglais n'ont pas le même sens ou la même force d'expression, certains termes seront conservés en anglais dans le texte. Cependant, le style de ces termes en anglais sera en *italique* dans le texte.

Ex. : Nous arrivons donc à la définition du *data warehouse.* Le *data warehouse* est l'entrepôt des données sur lequel les interrogations seront dirigées.

Partie 1
Introduction

L'introduction expose le cadre du projet dans son ensemble. Elle établit le contexte et la problématique de la prise de décision institutionnelle de l'UQTR. Seront présentés les principes de base des entrepôts de données ainsi que les questions et les réponses servant au raisonnement et à la logique de l'hypothèse de la solution proposée et ses limites. Elle se divise en trois chapitres, soit :

1. Introduction et problématique

2. Concepts de base

3. État des connaissances

«Nous espérons vivement que vous apprécierez la lecture de notre ouvrage et puisse celle-ci vous permettre de prendre des décisions éclairées pour la mise en place d'un système fonctionnel et efficace.»

Chapitre 1
Introduction et problématique

L'objectif du mémoire découle d'un besoin de support d'aide à la prise de décision. Pour atteindre cet objectif, un entrepôt de données sera construit pour offrir des tableaux de bord aux dirigeants. Ce chapitre permettra au lecteur de se familiariser avec les notions de base des entrepôts de données et de comprendre le contexte institutionnel de l'UQTR. Il est divisé en trois sections :

1.1 Introduction

Comment exploiter les données enfouies au creux des entrailles du système de gestion de base de données (SGBD). Dans les dernières années, les entreprises ont pris le virage technologique et ont accumulé des montagnes d'information, principalement dans leurs bases de données. Chaque processus d'affaires a été converti, passant du formulaire papier au formulaire en ligne. Les processus ainsi informatisés n'étaient pas nécessairement homogènes et on ne pouvait pas effectuer facilement le transfert, l'intégration ou le croisement des données entre eux. Aujourd'hui, le gestionnaire arrive à ce tournant : «Comment exploiter ces données emmagasinées afin d'aider à la prise de décision? ». Construire un entrepôt de données dans une perspective de support pour la prise de décision permettra, entre autres, d'offrir aux différents gestionnaires de l'Institution ce qu'il est commun d'appeler un tableau de bord.

> «Un tableau de bord est un instrument de mesure de la performance facilitant le pilotage «pro-actif» d'une ou de plusieurs activités dans le cadre d'une démarche de progrès. Le tableau de bord contribue à réduire l'incertitude et facilite la prise de risque inhérente à toute décision. Le tableau de bord est un instrument d'aide à la décision.» (Fernandez 05)

Si un gestionnaire souhaite obtenir la réponse à cette question : « Connaître le pourcentage d'étudiants diplômés dont le régime était à temps complet par rapport à ceux ayant un régime à temps partiel.», cette information n'est pas disponible directement dans les données textuelles, c'est-à-dire que l'information ne se retrouve pas dans le contenu d'un champ d'une seule table. Il faut effectuer des transformations, faire des calculs et des regroupements pour obtenir le résultat recherché. Ce problème pourrait se résoudre en utilisant des commandes SQL complexes. L'objectif de ce projet est de permettre au gestionnaire de poser ses questions et d'y trouver facilement ses réponses sans maîtriser le langage SQL ni être un expert des bases de données de l'entreprise. Pour atteindre cet objectif, un entrepôt de données sera construit.

Qui sont ces preneurs de décision? Il s'agit des décideurs (recteur, vice-recteurs, doyens, directeurs de services et de départements, et tout autre demandeur) désirant prendre une décision éclairée aussi simple ou complexe puisse-t-elle sembler. Les tableaux de bord s'adresseront à eux.

Dans les lignes qui suivent, afin de cerner la délimitation du projet, se retrouvent les sujets qui seront peu ou pas traités. Ce projet abordera brièvement les étapes de construction d'un tableau de bord. Là n'est pas le but premier. Il ne traitera pas d'archivage de données de l'entrepôt, ni d'application directe du «*data mining*». Il permettra seulement d'effleurer l'aspect ontologie des données. Ces trois derniers sujets feront partie intégrante des travaux futurs.

Parallèlement à ces travaux, voici les questions auxquelles on tentera d'apporter des réponses. Ces questions sont classées par ordre d'intégration temporelle.

Introduction du sujet

- Quelle est la différence entre une base de données et un entrepôt de données?
- Quelles sont les étapes de création de l'entrepôt institutionnel de données?

Phase de préparation des données

- Quelle est l'importance de la phase de préparation des données?
- Analyse des impacts temporels et événementiels :
    - Comment représenter les modifications des structures de données des systèmes de traitement de transactions en ligne (OLTP) afin de minimiser l'impact du côté de l'entrepôt de données et réduire les erreurs dans les résultats?
    - Comment informer le gestionnaire des événements majeurs pouvant affecter le résultat d'une analyse?

Phase de chargement

- Qu'est-ce que le processus d'extraction, de transformation et de chargement (ETC)?
- Quels logiciels existants peuvent répondre à nos différents besoins?

Présentation des données aux dirigeants

- Comment extraire les données pour les présenter aux dirigeants?

Il existe un vocabulaire propre au domaine (Tableau 1.1) qui sera expliqué plus amplement dans le chapitre 2 «Concepts de base». Cependant, pour faciliter la lecture immédiate, voici quelques-uns des principaux termes et leurs définitions.

Tableau 1.1
Vocabulaire spécifique au domaine

Terme Français (Anglais)	Définition
Entrepôt de données *(Data warehouse)*	Entrepôt des données (ED) sur lequel les interrogations sont dirigées.
Comptoir d'information *(Data mart)*	Magasin de données (MD) ciblé sur un ou plusieurs sujets. (Ex. : la facturation)
Cube de données *(Data cube)*	Cube de présentation d'unités pouvant pivoter sur différentes dimensions.
Forage de données *(Data mining)*	Technique de fouille des données pour en extraire de l'information complémentaire et des modèles de connaissance explicatifs ou prédictifs.
OLAP	*«Online Analytical Processing»* désigne les bases de données multidimensionnelles (aussi appelées cubes) destinées à des analyses complexes sur ses données.
ROLAP	Technique de modélisation et de stockage des données basée sur une structure relationnelle.
MOLAP	Technique OLAP optimisée pour l'analyse multidimensionnelle.
HOLAP	Technique hybride entre le MOLAP et le ROLAP.
Schéma relationnel	Schéma représenté par des entités et des relations (ER). Ce schéma est normalement en 3NF (3^e forme normale).
Schéma dimensionnel *(Dimensionnal modeling)*	Schéma représenté par des dimensions et des faits conformes. Ce schéma n'est pas en 3NF, on le dit «dénormalisé».
Schéma étoile *(Star schema)*	Schéma dimensionnel pouvant avoir une forme en étoile avec seulement une table de faits et plusieurs dimensions unitaires. La table de faits est placée au centre et les dimensions autour. C'est ce qui explique la forme en étoile.
Schéma en flocon *(Snowflake schema)*	Schéma dimensionnel pouvant avoir une forme en flocon de neige avec seulement une table de faits et plusieurs dimensions pouvant avoir une autre dimension en relation. La table de faits est placée au centre et les dimensions autour. Une dimension peut être associée à une autre dimension qui lui sera attachée. C'est ce qui explique la forme en flocon.

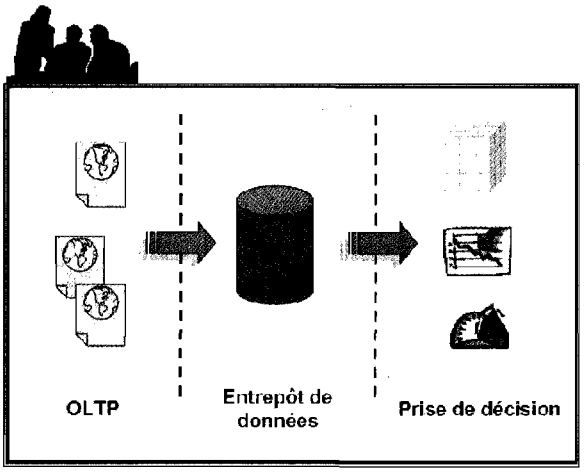


Figure 1.1 Vue d'ensemble de l'entrepôt.

Comme le montre la vue d'ensemble de l'entrepôt (Figure 1.1), l'intérêt du sujet sera plutôt situé en amont de l'entrepôt de données. Quelle méthodologie devra-t-on utiliser afin de charger les données des processus vers l'entrepôt? Quelles sont les étapes nécessaires? Comment combiner l'aspect temporel des données et des structures? Une fois ces questions résolues et réalisées, (ce travail représentant plus de 70% du processus total de réalisation de l'entrepôt), rendre accessible les données à la prise de décision sera presque un jeu d'enfant.

La problématique principale dans le contexte universitaire de l'UQTR est qu'il n'existe aucun soutien informatique facilitant la prise de décision. L'objectif premier est donc de mettre en place un prototype d'entrepôt de données dans l'environnement de l'UQTR et de proposer une méthodologie afin d'y parvenir. Pour ce faire, l'approfondissement du domaine est nécessaire. Il convient de définir une méthode de conception adaptée à l'UQTR et c'est pourquoi, l'étude des méthodologies de conception et l'analyse des outils existants est essentielle.

Une deuxième problématique, plus simple quoique essentielle, est de bien cerner les besoins des dirigeants. Ce sont eux qui maîtrisent leurs processus d'affaires. C'est à eux d'énoncer clairement leurs besoins. Le service informatique d'une organisation sécurise les données, transforme certains processus d'affaires afin de faciliter le travail des usagers, mais c'est aux dirigeants d'exprimer leurs besoins. Le processus d'affaires doit être connu des dirigeants avant que la demande au service informatique ne soit effectuée dans l'objectif d'une prise de décision. Le second objectif est d'aider les dirigeants à exprimer et à schématiser leurs processus d'affaires. Afin de simplifier cette phase, un questionnaire permettra d'aider les dirigeants à définir leurs processus d'affaires à l'intérieur de l'Université. Une fois ce schéma établi, il faudra cibler les données existantes en lien avec

le processus. Si toutes les informations essentielles sont disponibles, on pourra alors entreprendre le processus de création d'un premier comptoir de données.

Une autre problématique rencontrée se situe au niveau de l'intégrité des données. Dans certains systèmes, les clés étrangères ne sont pas au rendez-vous. La notion de clés étrangères est d'autant plus importante si elle est implantée dans les bases de données sources qui sont le point d'entrée de l'entrepôt et sauvent ainsi beaucoup de temps à la phase de «préparation des données». Pour transformer un schéma ER d'une base de données relationnelle vers un schéma dimensionnel, d'une base de données qui peut être relationnelle ou non, il faut «dénormaliser» le schéma ER c'est-à-dire passer de la $3^{ième}$ forme normale à la $2^{ième}$ ou même à la 1^{er} forme normale. Pour ce faire, il suffit de placer, par exemple, le libellé du nom du programme dans la table d'admission de l'étudiant au lieu du code du programme (qui peut être différent au numéro du programme utilisé dans les systèmes de l'Institution). Dans le schéma ER, le code du programme pointe sur une autre table, qui elle, contient le nom du programme. Si un code n'existe pas dans la table des programmes, il n'y aura aucun nom de programme à intégrer dans le schéma dimensionnel ce qui indique un manque d'intégrité des données. Le troisième objectif est de combler la lacune au niveau de la qualité et de l'intégrité des données. Nous avons créé quelques utilitaires informatiques sur mesure afin d'atteindre cet objectif. Des solutions seront proposées afin de corriger la situation.

Finalement, est survenue une autre problématique au cours du projet à laquelle il a fallu s'arrêter. Une problématique qui s'explique en deux points : Comment assurer l'exactitude des résultats aux dirigeants et comment informer les dirigeants des événements qui auraient pu influencer les données. Dans le jargon informatique, une traduction d'une expression anglaise serait : «Si les données d'entrée sont faussées, la sortie le sera tout autant...». C'est donc dans la phase de «préparation des données» qu'il fallait directement intervenir. Le dernier objectif s'énonce alors comme suit :

> «Comment assurer l'exactitude des résultats aux dirigeants» se traduit par le fait qu'il faut gérer l'évolution des structures de données dans le temps et ainsi en minimiser l'impact sur l'entrepôt de données. Si nous ajoutons un champ en 2003 et que nous analysons ce champ sur les années 2000 à 2005, il y aura une absence

de données dans la période précédant l'ajout du champ. Il y a plusieurs façons de résoudre ce problème, mettre une des valeurs suivantes dans le champ : «NULL[1]», 0, -1, NA, 'inconnu'. Comment sera interprétée une absence de valeur pour une année où le champ existe vraiment? Si on fait une moyenne ou une somme des valeurs agrégées par année, en perdant le détail de chaque ligne pour l'année, l'année 2005 qui chevauche la création du champ aura des valeurs inexactes. Du même principe, la modification et la suppression causeront de l'incertitude dans les résultats. C'est pourquoi on voudra informer le gestionnaire sur l'historique du champ qui a été requis pour obtenir un résultat.

«Comment informer les dirigeants des événements qui auraient pu influencer les données» se traduit plutôt par un fait ou un événement dans le temps qui influence le cours normal des processus universitaires. Par exemple, la grève des chargés de cours qui a fait augmenter le taux d'échec ou la cote «incomplet» à une session donnée. C'est pourquoi, privé de cette information, le gestionnaire demeure sans explication quant à un fait empirique mesurable qui influence négativement, par exemple, la moyenne et l'écart-type.

Cela dit, voici donc le résumé des objectifs à atteindre :

- Créer un prototype d'entrepôt institutionnel de données et proposer une méthodologie de conception.
- Guider les dirigeants à schématiser leurs processus d'affaires.
- S'assurer de l'intégrité référentielle des systèmes sources.
- S'assurer de la fiabilité et de l'exactitude des résultats en informant les dirigeants.

[1] NULL : Dans les bases de données, un champ vide qui ne contient aucune valeur est représenté par la valeur explicite «NULL»

1.2 Les systèmes opérationnels de l'UQTR

L'UQTR possède plus de 70 systèmes transactionnels de traitement de données en ligne (OLTP). Quelques systèmes sous PowerHouse sont actuellement en conversion vers le Web, mais dans les deux cas, les données reposent sur une même base de données Oracle. L'analyse croisée des données des différents systèmes prend un temps ressource (homme/machine) énorme. De plus, une connaissance d'expert est requise afin de trouver ces données, et ce, à des fins d'analyse, d'interprétation et d'exploration. Les données d'analyse dorment donc au creux des entrailles du SGBD et ne sont pas accessibles facilement pour la prise de décision. La création d'un entrepôt de données est donc essentielle pour la construction des tableaux de bord demandés par la haute direction et permettra ainsi d'offrir une accessibilité et une compréhension intuitive des données pour les demandeurs.

Voici un extrait du discours du Recteur de l'UQTR du 18 octobre 2007 exposant le besoin pressant de la mise en place de l'entrepôt institutionnel de données pour les tableaux de bord :

> *«Depuis l'épisode des contrats de performance, les universités au Québec sont assujetties à une batterie de vérifications, toutes redevables des redditions de compte bardées d'objectifs, de paramètres, d'indicateurs et de cibles. À telle enseigne que nos décisions se trouvent associées à des contraintes imposées de l'extérieur de l'Université, échappant ainsi au contrôle interne et, plus spécifiquement, à l'analyse globale de nos instances. C'est pourquoi il m'apparaît pressant de doter l'UQTR d'un outil capable d'aider au travail de gestion. Un outil qui soit en mesure de regrouper les informations essentielles au bon fonctionnement de chacune des unités de notre institution, et qui permette à la fois de suivre l'évolution de ces unités et de fournir les arguments autorisant un développement stratégique.*
>
> *Cet outil, apte à favoriser la gestion et le développement organisationnels, devra être suffisamment souple pour permettre l'application de fonctions-contrôle et, de même, pour détecter les opportunités qui se présentent aux fins*

> *de notre croissance. Soyons clairs, il ne s'agit pas d'un outil de strict contrôle, mais bien d'un système qui permettra à notre institution de prendre des décisions éclairées.*
>
> *Vous aurez compris qu'il est ici question d'une opération qui porte le nom de « Tableau de bord ». »*

Si un gestionnaire veut obtenir des informations à des fins d'analyse, il se retrouve actuellement dans l'impossibilité de cheminer de manière autonome. Il demande donc une extraction de données au service informatique qui lui retourne un fichier statique Excel sans plus. Il existe aussi un logiciel qui permet à un gestionnaire détenant une connaissance des tables de la base de données d'être plus autonome. Ce logiciel est «Discoverer Web». Il a ses limites et sans une connaissance exacte de l'information enfouie dans les tables, le gestionnaire éprouve alors d'énormes difficultés d'extraction et de traitement des données.

Un recensement de l'inventaire complet de tous les systèmes de l'UQTR a été minutieusement réalisé homologuant le nom des systèmes, leur code, leur description ainsi que le nom du responsable. Chaque système a été analysé et catégorisé. Il est essentiel de déterminer les systèmes de type «services à la communauté » car ces systèmes sont à exclure des systèmes susceptibles d'être retenus pour la prise de décision. Parmi les 70 systèmes existants, seulement une vingtaine d'entre eux seront retenus à des fins d'analyse. Dans l'inventaire, d'autres informations y sont présentées. On y retrouve la base de données sur laquelle repose le système, et l'information qui nous dit si le système est accessible en ligne ou encore sur l'ancienne plate-forme Power House. Un tableau des systèmes gérés par le développement des systèmes d'information (DSI) de l'UQTR a été placé en annexe en considérant l'ampleur de l'information contenue. Le lecteur pourra le consulter à l'annexe A.

Suite à cet inventaire, chaque table a été examinée. Dans le schéma opérationnel, plus de 1550 tables reposant sur une base de données relationnelles Oracle sont représentées pour l'ensemble des systèmes de gestion de l'UQTR. La répartition des données emmagasinées dans trois zones distinctes représente approximativement 20 giga-octets

(Go) d'espace-disque. En examinant le tableau 1.2 de la répartition physique des données des systèmes, on peut estimer une augmentation moyenne de volume de 35% annuellement pour les systèmes transactionnels. Cette taille sera directement représentative de l'ampleur de l'entrepôt de données.

Tableau 1.2
Répartition physique des données

Zone de stockage des données	Volume sept 2006	Volume sept. 2007	Augmentation	Pourcentage
Audit-Web	4.9 Go	9 Go	4.1 Go	45.5%
DAF (Dossier étudiants)	1 Go	2.5 Go	1.5 Go	60%
Tous les autres	6.8 Go	8 Go	1.2 Go	15%
Total :	12.7 Go	19.5 Go	6.8 Go	34.87%

Dans un premier temps, une analyse plus poussée a été réalisée relativement aux tables en rapport avec l'admission d'un candidat. Une constatation en découle : il n'y a pas de clé étrangère activée dans la plupart des tables qui sont en relation, ce qui complexifie la phase de préparation des données. Ce point sera abordé au chapitre 6 au niveau de la conception.

Proposition d'un cas type de tableau de bord
Par Rémy Auclair

Agent de recherche
Service des affaires départementales de l'UQTR

Les trois paragraphes qui suivent, composés par Rémy Auclair, décrivent une demande type de tableau de bord. Sera réalisé dans le cadre du projet, le suivi des étapes du processus d'admission correspondant au 2e paragraphe.

«La direction des affaires départementales, plus précisément son secteur de la recherche institutionnelle, produit, d'une part, des données institutionnelles pour alimenter les réflexions continues sur l'orientation et les choix stratégiques et, d'autre part, des travaux d'analyse dans le but d'aider et de soutenir le processus décisionnel des unités académiques et administratives de l'Université. Le présent modèle de tableau de bord viserait à consolider la capacité d'intervention de l'Université en matière de suivi de clientèles étudiantes et de permettre aux divers intervenants d'ajuster leurs activités à la lumière d'une connaissance fine de l'évaluation de ces mêmes clientèles.»

«Premièrement, il devrait permettre de suivre les diverses étapes du processus d'admission à un trimestre donné. Le tableau de bord ainsi constitué permettrait de suivre progressivement les étudiants à partir de l'étape de la demande d'admission en passant par les candidats acceptés, les nouveaux inscrits (dans l'établissement et dans le programme) et les inscriptions totales. Le tout serait disponible selon un certain nombre de caractéristiques : le programme, le genre de programme, le sexe, le régime d'études, le groupe d'âge, le collège et la région de provenance, la CRC (cote R au collégial) et le diplôme comme base d'admission.»

«Deuxièmement, il devrait également permettre de suivre progressivement le cheminement des étudiants (inscrits dans le même programme ou dans un autre programme) jusqu'à leur destin scolaire, soit la diplômation ou son contraire, l'abandon. Il serait également intéressant de jumeler les données des enquêtes ICOPE (lorsqu'elles seraient disponibles) au cheminement des étudiants de manière à calculer, par régression logique, des facteurs favorisant la réussite de leurs études.»

1.3 Plan et contenu

Le document est divisé en trois (3) parties. La première, la partie introductive, est celle qui introduira le sujet, le contexte, les concepts de base et la problématique. À la lecture de la partie introductive, le lecteur aura une idée précise de l'ensemble du contexte et de la problématique. Cette première partie contient trois (3) chapitres : «Introduction et problématique»; «Concepts de base» et «État des connaissances». La seconde partie est la méthodologie. Elle posera les bases de la méthodologie proposée. L'analyse et la justification des choix y seront expliquées. De plus, une étude des outils existants a été réalisée menant à une recommandation d'achat pour l'UQTR. Cette deuxième partie contient deux (2) chapitres : «Analyse» et «Études des outils logiciels existants». La troisième partie «Conception de l'entrepôt de données», se compose de deux (2) chapitres : «Conception du modèle» et «Publication des données». La création du comptoir de données de l'admission sera effectuée en suivant cette méthodologie. Une fois réalisés, quelques tableaux de bord seront construits et présentés. Enfin, une conclusion terminera le document en y présentant, notamment, quelques hypothèses de développements ultérieurs.

Chapitre 2
Concepts de base

Dans ce chapitre, nous aborderons les notions de base essentielles à la poursuite du document. Ce chapitre permettra au lecteur de connaître davantage les indicateurs de performance décisionnelle et de comprendre les différences entre les bases de données relationnelles et les entrepôts de données. C'est une introduction nécessaire au domaine avant d'aborder le chapitre sur l'état de l'art qui lui est plus technique dans son argumentation. Il est divisé en deux sections :

2.1 Le décisionnel

2.2 Base de données relationnelle et entrepôt de données

2.1 Le décisionnel

Décider, c'est prendre des risques certes mais c'est être en réaction face à un choix stratégique à prendre.

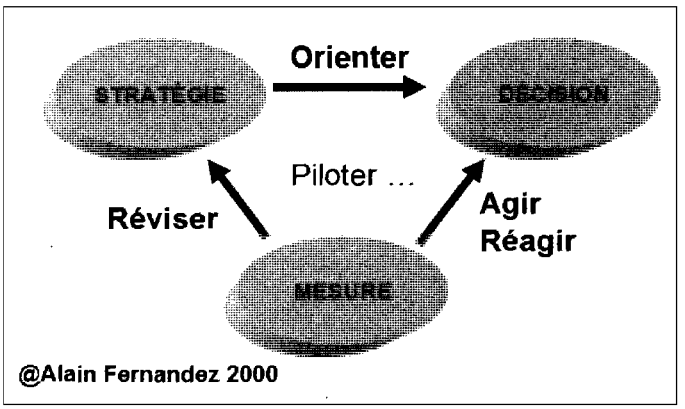


Figure 2.1 Lien entre les mesures, la stratégie et la décision.
Source : [Fernandez 05]

La figure 2.1 parle d'elle-même. Selon Alain Fernandez, pour prendre une décision, il faut pouvoir mesurer. La mesure permettra de réviser la stratégie pour orienter la décision et ainsi d'agir ou de réagir. La mesure représente l'indicateur qui permet d'évaluer la performance d'où le terme «indicateur de performance». De nos jours, cette mesure n'est pas uniquement d'ordre financier, comparativement au passé dans la littérature. Ces indicateurs, représentés sous forme de tableau de bord, sont nécessaires à tous les niveaux hiérarchiques comme le montre la figure 2.2.

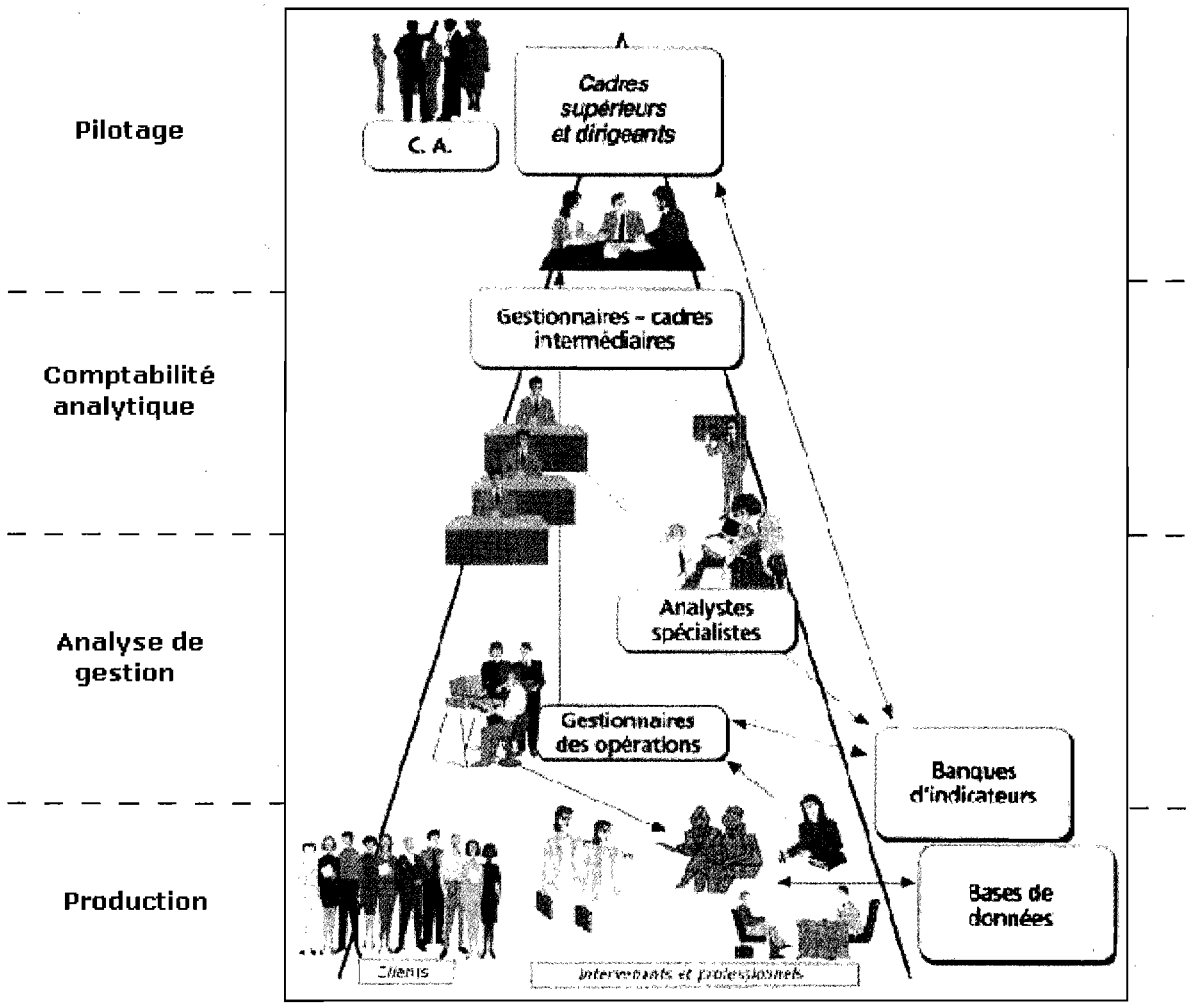


Figure 2.2 Les tableaux de bord et la structure hiérarchique.
Source : [Voyer 02]

Les différents niveaux de l'entreprise ne traitent pas des mêmes besoins, mais les données se recoupent nécessairement. À la base de la pyramide, les clients et les intervenants correspondent au niveau «production». C'est à ce niveau que les données sont emmagasinées dans les bases de données. Le niveau supérieur se nomme «analyse de gestion» et représente les gestionnaires des opérations et les analystes spécialistes. L'avant-dernier niveau se nomme «comptabilité analytique» et couvre surtout les gestionnaires et cadres intermédiaires. Finalement, dans le haut de la pyramide au niveau «pilotage» on y retrouve les cadres supérieurs et les dirigeants. Chaque niveau doit prendre ses propres décisions en fonction des objectifs globaux de l'entreprise.

Selon Wikipédia, *«la prise de décision est un processus cognitif complexe visant à la sélection d'un type d'action parmi différentes alternatives. Il s'agit d'une méthode de raisonnement pouvant s'appuyer sur des arguments rationnels et/ou irrationnels. Il existe différents niveaux de décision qui doivent être pris dans une entreprise : les décisions stratégiques, les décisions tactiques et les décisions opérationnelles.»*

Tableau 2.1
Types de décision selon Wikipédia

Caractéristiques	Stratégique	Administrative	Opérationnelle
Domaine de la décision	Relations avec l'environnement	Gestion des ressources	Utilisation des ressources dans le processus de transformation
Horizon de temps	Moyen et long terme	Court terme	Très court terme
Effet de la décision	Durable	Bref	Très bref
Réversibilité de la décision	Nulle	Faible	Forte
Procédure de décision	Non programmable	Semi programmable	Programmable
Niveau de la prise de décision	Direction générale	Directions fonctionnelles	Chefs de services ou d'atelier
Nature des informations	Incertaines et exogènes	Presque complètes et endogènes	Complètes et endogènes

Selon le tableau 2.1, chaque type de décision sous-entend des caractéristiques les définissant mais aussi limitant leurs champs d'action dans le temps. Il ne faut pas seulement vouloir prendre une décision, il s'agit de la prendre en étant le mieux éclairé possible. De nos jours, les tableaux de bord avec indicateur de performance peuvent répondre efficacement à la prise de décision. Cependant, il ne s'agit pas simplement de vouloir des tableaux de bord, il faut savoir ce que l'on veut y retrouver.

À quoi sert un tableau de bord et comment se construit-il ? Selon [Fernandez 2005], les cinq rôles des tableaux de bord se décrivent comme suit : réduire l'incertitude; stabiliser l'information; contribuer à une meilleure maîtrise du risque; faciliter la communication et dynamiser la réflexion. Il [Fernandez 05] propose une méthode efficace en cinq étapes et quatorze outils pour construire ces tableaux de bord de gestion. Le lecteur pourra consulter en annexe B le détail de la méthodologie de Fernandez basée sur la méthode Gimsi.

- Étape 1 : sélectionner les axes de progrès (sujets, attentes, leviers, progrès).
- Étape 2 : déterminer les points d'intervention (processus et activités critiques).
- Étape 3 : choisir les objectifs (choix, risques, plan d'action).
- Étape 4 : choisir et construire les indicateurs.
- Étape 5 : bâtir et maintenir le tableau de bord.

Les titres des étapes suffisent quant à eux à guider l'usager afin de bien définir ses besoins avant de faire une demande de tableau de bord auprès des instances concernées.

Quel est le lien entre la prise de décision, les tableaux de bord et l'entrepôt de données ? Ils sont en étroite symbiose. L'un est nécessaire à l'autre, ils forment un tout. Le tableau de bord est un outil qui aide à la prise de décision. Il est le moyen d'expression des résultats permettant de prendre action. Le tableau de bord présente des informations qui sont quant à elles structurées dans l'entrepôt de données. L'entrepôt de données permet d'offrir performance et facilité d'interaction par l'usager lui-même. Avant d'offrir des tableaux de bord aux usagers, il faut préparer les données afin de les intégrer à l'entrepôt. Pour pouvoir offrir une information, il faut nécessairement avoir la donnée correspondante emmagasinée quelque part dans l'entreprise, sinon on ne peut l'inventer.

2.2 Base de données relationnelle et entrepôt de données

Quelle est la différence entre une base de données relationnelle traitant des transactions journalières et un entrepôt de données? Le lecteur connaîtra ici la réponse à cette question.

Les systèmes de traitement de transactions en ligne (OLTP), aussi appelés systèmes transactionnels ou opérationnels, permettent d'effectuer les saisies, les enregistrements et les mises à jour des transactions journalières d'un système de gestion quelconque. Ces données sont conservées majoritairement dans une base de données relationnelle. Une base de données relationnelle s'assure de l'intégrité des données et repose sur un schéma entité-relation (ou son équivalent ER). Le schéma ER se transpose en structure relationnelle normalisée souvent à la 3^{e} forme normale (3NF) de telle sorte que la répétition des données est à son minimum. La figure 2.3 schématise, par des entités («GEN_PERSONNE», «DAF_DICTIONNAIRE»,...) et des relations, la structure de l'admission de l'UQTR (Figure 2.3). La table «GEN_PERSONNE» identifiée toute personne en lien avec l'UQTR : employés, étudiants et membres du centre sportif. La table «DAF_DICTIONNAIRE» représente les personnes qui sont étudiants à l'UQTR. Lorsqu'un étudiant fait une demande d'admission un ajout se fait dans la table «DAF_ETUD_CAND» indiquant quel étudiant a fait une demande d'admission pour quel programme. La table «TRI_PROGRAMME» identifie les programmes de l'UQTR. Une demande d'admission est associée à une base d'admission selon le cycle et le programme. La table «DAF_TB_BADM» contient les bases d'admissions possibles pour tous les étudiants. La table «DAF_BASE_ADMI» contient les bases d'admission associées à un programme pour un étudiant qui fait une demande d'admission.

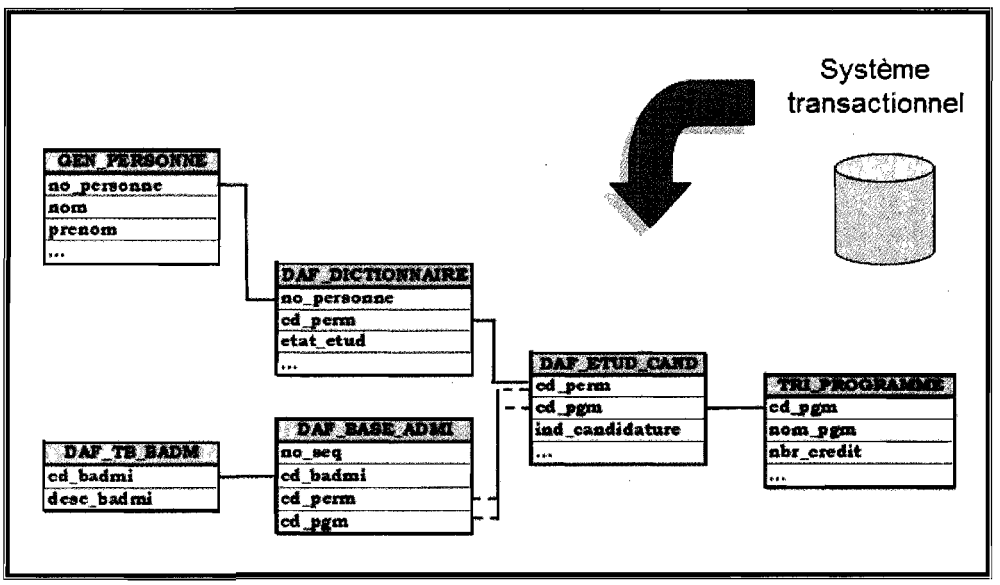


Figure 2.3 Schéma entité-relation(ER) de l'admission UQTR.

Plus la structure est normalisée, plus l'analyse des données sera complexe et lente. Il y aura un grand nombre de tables et donc de jointures, ce qui entraîne une diminution des performances, une complexité des requêtes pour obtenir l'information souhaitée, et un temps de traitement plus long.

L'aspect sécurité est présent dans la plupart de ces systèmes ER. Pensons par exemple aux guichets automatiques des institutions bancaires ou à «Accès D» de Desjardins permettant aux membres d'effectuer des opérations sur leur compte de caisse en ligne avec une identification et un mot de passe.

Les systèmes transactionnels sont conçus dans différentes bases de données de l'entreprise et peuvent être dispersés dans des services distincts. Il devient alors impossible de faire du recoupement de données à des fins d'analyse puisqu'ils ne sont pas construits dans ce but précis.

Retenons les principales caractéristiques des systèmes transactionnels :

- Rapidité à mettre à jour un enregistrement précis.
- Assure l'intégrité des données.
- Minimise la redondance des données.
- Systèmes en vase clos, associés à un processus d'affaires (Un système pour la vente, un système pour la comptabilité, un système pour la production, etc.). Ces systèmes ne peuvent pas nécessairement échanger des données entre eux.
- Structure relationnelle.

Pour les entrepôts de données, nous introduisons immédiatement la définition suivante du livre d'Inmon [Inmon 96]:

> *« Un entrepôt de données est une collection de données portant sur des sujets touchant une organisation, intégrées, variant dans le temps, et non-volatiles pour supporter le processus de prise de décision d'une organisation »*

Les entrepôts de données permettent de centraliser les informations stratégiques des différents systèmes de gestion. Ils s'alimentent des systèmes transactionnels. Tous les systèmes qui nécessitent un traitement décisionnel sont intégrés à l'entrepôt, ce qui permet une analyse croisée sur l'ensemble des données de l'entreprise. Les données pouvant être agrégées sont chargées distinctement des données journalières, ce qui augmente la performance du temps de réponse. Par exemple, conserver les ventes mensuelles dans une table et les transactions de ces ventes dans une autre table. Si l'usager veut analyser les ventes mensuelles des trois dernières années, puisque les valeurs n'ont pas besoin d'être calculées, une seule table sera accédée.

Une autre caractéristique distinctive des entrepôts de données est le fait de conserver les données de façon historique. Ce principe s'explique comme suit : «chaque modification d'enregistrement est conservée distinctement des valeurs précédentes avec une valeur temporelle qui situe cette modification dans le temps ». Ce qui veut dire que l'on peut ajouter des données à l'entrepôt sans écraser les anciennes valeurs. Si un étudiant déménage cinq fois durant ses études, le système transactionnel conservera seulement la dernière adresse de l'étudiant. L'entrepôt conservera les cinq adresses différentes ainsi que la date des changements d'adresse. Ce phénomène permettra d'analyser dans le temps la provenance des étudiants à un moment précis. L'adresse de l'étudiant sera l'adresse valide au temps *t* choisi pour l'analyse. On retrouve dans l'entrepôt les données actuelles, présentement «vraies» dans les systèmes transactionnels, mais en plus les archives temporelles de ces données.

Les entrepôts de données peuvent reposer sur les bases de données relationnelles, mais le schéma utilisé n'est pas un schéma ER. Un schéma dimensionnel dans lequel les tables sont structurées de façons différentes permet une performance optimale et une meilleure compréhension de la part de l'usager. La figure 2.4 illustre bien le schéma dimensionnel de l'admission de l'UQTR. Les tables du schéma de la figure 2.3 ont été dénormalisées et structurées différemment. La table de faits «FACT_ADMISSION», centrale au schéma en étoile, représente les demandes d'admission. Trois dimensions sont associées à cette table de faits : les étudiants «DIM_ETUDIANT», les programmes «DIM_PROGRAMME» et le temps «DIM_TEMPS». Nous aborderons la notion temporelle de la table «DIM_TEMPS» au point 3.6 du chapitre 3.

Le lecteur constatera que le schéma de la figure 2.4 est plus intuitif pour l'utilisateur que le schéma ER de la figure 2.3. Sémantiquement, dans la figure 2.4, une table représente un concept qu'il est facile de comprendre. L'utilisateur pourra retrouver plus facilement les données qu'il recherche. Puisqu'il y a moins de tables, il y aura moins de jointures et l'accès aux données sera plus rapide.

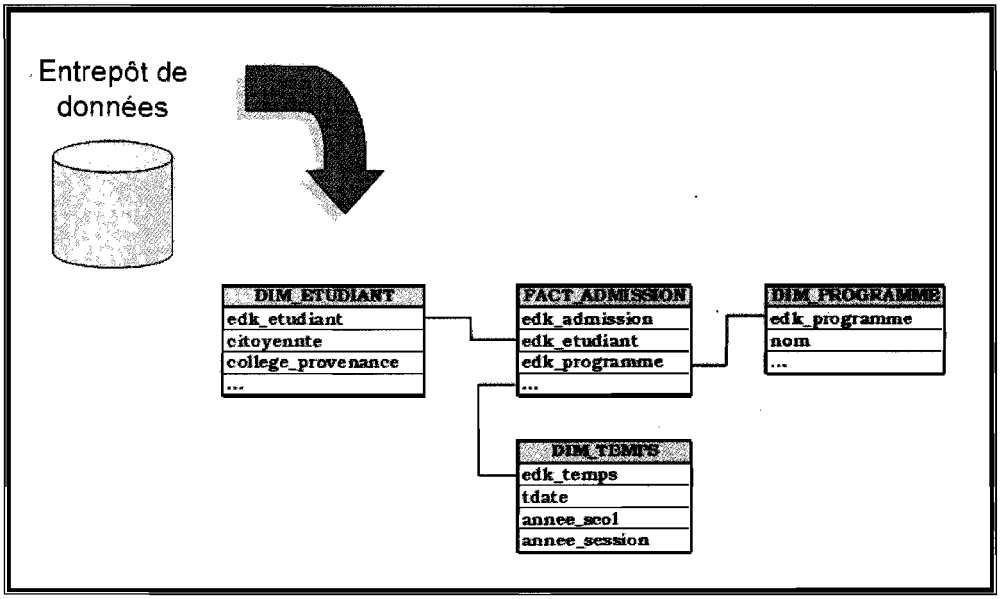


Figure 2.4 Schéma dimensionnel de l'admission de l'UQTR.

Retenons les principales caractéristiques des entrepôts de données :

- Centralisation des données des différents systèmes de l'organisation.
- Données récapitulatives (agrégées, résumées ou sommaire).
- Données historiques (actuelles et d'archives).
- Grand volume de données.
- Pas de mise à jour, uniquement des ajouts de données.
- Performance.
- Structure dimensionnelle.

Chapitre 3
État des connaissances

Dans ce chapitre, les bases du domaine sont recensées afin d'obtenir une vue d'ensemble des sujets touchant la mise en place d'un entrepôt de données. Lors de notre recension de la littérature sur le sujet, notre choix des articles s'est justifié en fonction de la diversité et de l'importance des différents sujets à traiter autour des entrepôts de données. Ce chapitre est donc divisé en dix (10) sections :

3.1 Introduction aux entrepôts de données

Dans le monde des affaires, toutes sortes d'information sont emmagasinées dans différents systèmes de l'entreprise. Ces données sont récupérées par différents outils afin d'être analysées par un expert humain, qui après traitement semi-automatique ou non, prendra une décision. Cette décision aura un impact positif ou négatif affectant l'entreprise à court ou à moyen terme.

Un entrepôt est un ensemble de données de différentes sources de l'entreprise centralisé et homogénéisé. Les systèmes ciblés pour des fins d'analyse seront incorporés à l'entrepôt pour ne former qu'un seul et unique système et ainsi d'aider les preneurs de décision en leur permettant d'accéder rapidement à l'information via des outils adaptés.

Selon [SEN-SINHA 05], un entrepôt de données est : orienté-sujet, intégré, variant dans le temps et non volatile. Cette définition est dérivée de celle du créateur du terme «entrepôt de données» monsieur Bill Inmon [Inmon 96]. Un entrepôt de données «orienté-sujet» est compartimenté en sujets principaux, représentant les processus d'affaires majeurs. Ce fractionnement permet l'analyse transversale entre les processus de l'entreprise et empêche la répétition des sujets.

Avant d'être intégrées à l'entrepôt, les données devront être analysées, nettoyées et transformées. Les données qui entrent dans l'entrepôt se doivent d'être intègres et exactes pour ne pas fausser les résultats des analyses lorsqu'elles seront exploitées par les logiciels de présentation. L'entrepôt de données permet de consolider deux types de données : les données détaillées et les données agrégées. Les données détaillées représentent les transactions courantes qui doivent être chargées quotidiennement dans l'entrepôt. Leur granularité de chargement est journalière. Les données agrégées correspondent aux besoins d'analyse des utilisateurs. La fréquence de chargement dépend du niveau de granularité qui peut être mensuel, trimestriel ou annuel selon le besoin d'analyse requis.

Lorsqu'une donnée entre dans l'entrepôt, selon sa fréquence de changement, elle ne doit jamais être modifiée, et ce, afin de fournir exactement le même résultat à une requête à chaque fois qu'elle est requise pour la même période donnée. Puisque les décisions sont prises en fonction du résultat obtenu, une donnée ne peut être ni modifiée, ni supprimée dans le passé. C'est pourquoi on dit que les données sont non volatiles dans un entrepôt.

Puisque les données sont non volatiles, il faut pouvoir retracer dans le temps la valeur exacte d'un fait au moment désiré. Des indicateurs temporels sont associés aux données afin de récupérer la valeur valide au moment t voulu.

Selon Inmon [Inmon 96] et Lawyer [LAWER-CHOWDHURY 07], il y a deux (2) propriétés essentielles d'un entrepôt, la flexibilité et l'extensibilité. Les questions (besoins) des utilisateurs d'aujourd'hui ne seront pas les mêmes demain. Il faut donc prévoir par la flexibilité la modification des structures et des données dans le temps. Il faut aussi prévoir l'augmentation du volume des données, l'augmentation du nombre d'utilisateurs et la complexité des requêtes usager.

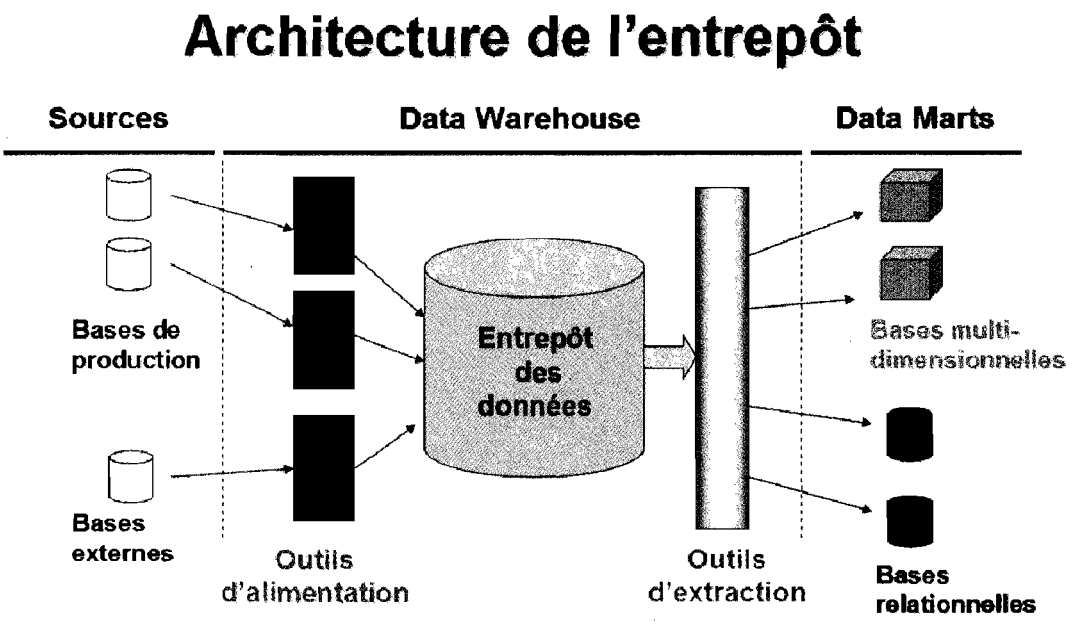


Figure 3.1 Vue détaillée de l'architecture de l'entrepôt.

La figure 3.1 présente la vue détaillée de l'architecture de l'entrepôt. Les sources seront intégrées à l'entrepôt par des outils d'alimentation. Les tableaux de bord seront produits après qu'une demande d'information autorisée soit présentée aux outils d'extraction.

3.2 Approches générales

D'après [LIST et al. 02], construire un entrepôt de données est un défi de taille. Il existe plusieurs approches de modélisation mais trois approches sont communes à [LIST et al. 02], [TDWI 04] et [SHI et al. 01]. Il s'agit de l'approche «top-down», de l'approche «bottom-up» et de l'approche «hybride» qui est un mélange des deux premières approches. Le tableau 3.1 présente les caractéristiques des approches de base.

Tableau 3.1
Caractéristiques des approches de base à la construction d'un entrepôt de donnée
Source : http://www.systemeETC.com/approches_dw.htm

Top-Down (Bill Inmon et le CIF)	Bottom-Up (Ralph Kimball et le Bus Architecture)	Hybride
Caractéristiques majeures		
➢ L'emphase est mise sur le DW. ➢ Commence par concevoir un modèle de DW au niveau de l'entreprise. ➢ Déploie une architecture multi tiers composée de staging area, le DW, et les data mart dépendants. ➢ Le staging area est permanent. ➢ Le DW est orienté entreprise; les data marts sont orientés processus. ➢ Le DW contient des données atomiques; Les data marts contiennent les données agrégées. ➢ Le DW utilise un modèle de données normalisé de toute l'entreprise; Les data marts utilise des modèles dimensionnels orientés sujet. ➢ Les utilisateurs peuvent effectuer des requêtes sur le DW et les data marts.	➢ L'emphase est mise sur les data marts. ➢ Commence par concevoir un modèle dimensionnel pour le data mart. ➢ Utilise une architecture qui consiste en un staging area et les data marts. ➢ Le staging area est en général non permanent, mais il peut devenir permanent pour implanter l'architecture en BUS (Dimensions et faits conformes) ➢ Les data marts contiennent les données atomiques et les données agrégées. ➢ Les data marts peuvent fournir une vue entreprise ou processus. ➢ Un data mart consiste en un seul schéma étoile physique. ➢ Les data marts sont implantés d'une façon incrémentale et intégrée en utilisant les dimensions conformes. ➢ Les utilisateurs ne peuvent effectuer des requêtes sur le staging area.	➢ L'emphase est sur le DW et les data marts; utilise les deux approches "top-down" et "bottom-up" . ➢ Commence par concevoir un modèle de données de l'entreprise en même temps que les modèles spécifiques. ➢ Passe 2–3 semaines à créer un modèle normalisé d'entreprise de haut niveau ; génère les modèles des premiers data marts. ➢ Charge les data marts avec les données atomiques en utilisant un staging area temporaire. ➢ Les modèles des data marts sont composés d'un ou plusieurs star schemas. ➢ Utilise un outil ETC pour charger les data marts et pour échanger le métadata avec ces derniers. ➢ Charge le DW à partir des data marts lorsqu'il y a besoin de faire des requêtes à travers plusieurs data marts en même temps.

De ces approches de base, les auteurs [LIST et al. 02] ont dérivées trois (3) approches orientées soient par les données, les utilisateurs ou les objectifs : l'approche «piloter par les données», l'approche «piloter par les besoins des utilisateurs» et l'approche «piloter par les objectifs».

L'approche «piloter par les données», associée à Inmon, est basée sur le schéma ER des systèmes transactionnels. Toutes les données doivent être chargées sans nécessité de connaître a priori les besoins des utilisateurs.

L'approche «piloter par les besoins utilisateurs» [WINTER-STRAUCH 02], associée à Kimball, est l'approche stratégique utilisée par les magasins Wall-Mart. Cette approche s'assure que les buts de tous et chacun vont dans la même direction. Les utilisateurs priorisent leurs besoins qui eux, en fonction des besoins plus généraux à l'intérieur de l'entreprise, sont priorisés de nouveau dans un ensemble global. Les utilisateurs sont sollicités afin de mieux cerner les processus d'affaires.

L'approche «piloter par les objectifs» [GIORGINI et al. 05], associée à leur processus de modélisation SOM (Semantic Object Model), détermine en premier lieu les objectifs et les services de l'entreprise existants. Par la suite, une analyse des interactions entre les clients et les transactions pour un processus donné est réalisée. L'étape suivante est de convertir les séquences transactionnelles pour trouver la dépendance des séquences. Finalement, les mesures et les dimensions sont identifiées. Cette approche est efficace seulement si les processus sont en lien avec les objectifs de l'entreprise.

Tableau 3.2
Comparaison des approches orientées pour la conception des entrepôts
[LIST et al. 02]

Méthodologie Critères	Piloter par les données	Piloter par les besoins usagers	Piloter par les objectifs
Approche de base	Bottom-up	Bottom-up	Top-Down
Auteur	William(Bill) Inmon	Ralph Kimball	Böhnlein and Ulbrich-vom Ende
Méthode directive	TCS (Taylorism Classical School)	Aucune (reflet de la culture de l'entreprise)	Par objectif
Supporté par	Utilisateurs	Départements	Haut direction
Niveau organisationnel cible	Opérationnel	Dépend du groupe d'intervenant	Stratégique, Administratif, et opérationnel
Nombre de mesures	Plusieurs	Plusieurs	Peu
Nombre de dimensions	Peu	Plusieurs	Peu
Nombre de systèmes sources	Faible	Moyen	Élevé
Longévité et stabilité du modèle de données	Long	Court	Long
Coût	Faible	Élevé	Élevé

Le tableau 3.2 présente brièvement quelques critères des approches générales. Les auteurs concluent que la méthode «piloter par les besoins des utilisateurs» est risquée et doit être combinée avec l'une des deux autres méthodes. L'approche «piloter par les objectifs» est plus directement associée au cube de données et, donc, à la prise de décision, et que l'approche «piloter par les données» s'applique plus particulièrement au forage de données.

3.3 Rôles de l'entrepôt de données

Les données opérationnelles sont définies dans les processus organisationnels de l'entreprise. Chaque processus possède ses données. Les données ne sont pas nécessairement les mêmes entre les processus et elles ne représentent pas les mêmes éléments.

Les systèmes d'aide à la décision à des fins d'analyse ont besoin des informations de tous les processus de l'entreprise. Ils sont orientés-sujets ou orientés-processus.

L'entrepôt permettra d'intégrer les différents systèmes opérationnels et de trouver les liens possibles entre les systèmes afin de permettre l'analyse croisée entre les différentes fonctions de l'entreprise et ce, par une méthode de «RÉINGÉNIERIE». L'entrepôt ne sera en aucun cas une copie des systèmes opérationnels de l'entreprise (OPTL).

L'entrepôt permet de faciliter la prise de décision. Les données sont intégrées vers l'entrepôt et sont extraites afin d'aider les gestionnaires à poser une action. Comment les données des systèmes transactionnels cheminent-elles vers la prise de décision ?

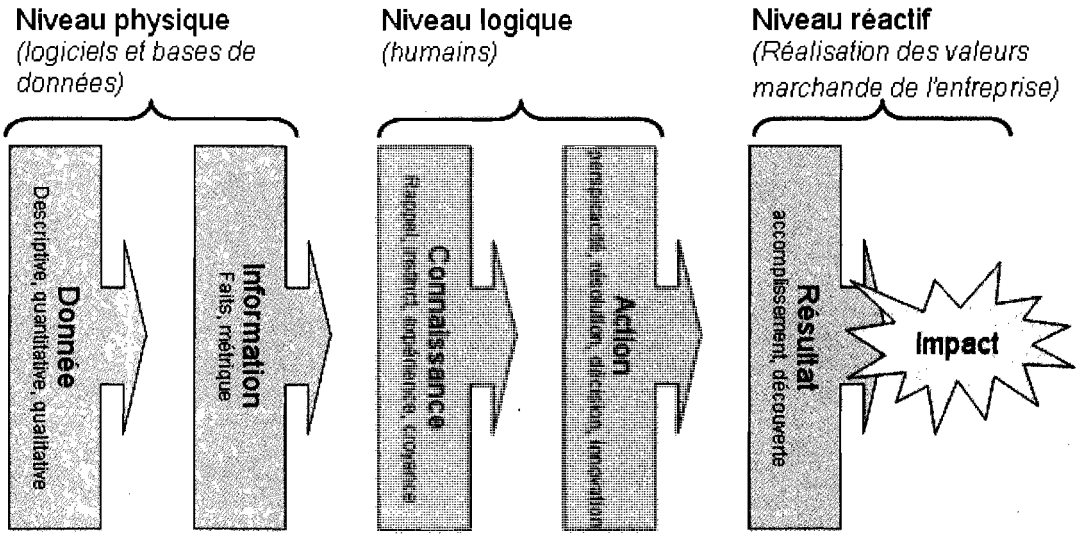


Figure 3.2 Donnée, information et connaissance. [TDWI 04]

Le livre blanc des concepts et principes des entrepôts de données [TDWI 04] représente le cheminement comme étant une transformation des données en information vers les connaissances. Ce processus (figure 3.2) est catégorisé en trois niveaux hiérarchiques permettant de réaliser différents objectifs.

Les données sont soient des données opérationnelles ou des données informationnelles intégrées et nettoyées afin de constituer la matière première avec laquelle l'information est construite. L'information est une collection organisée de données dans un contexte intelligible précis afin d'informer les personnes et les processus de faits et de métriques.

La connaissance est un ensemble d'informations sur un sujet donné qui forme une nouvelle connaissance pour la personne qui observe en fonction de facteurs prédéfinis par lui-même. Par la suite une action peut être posée dans le «bon sens» après analyse des connaissances.

Les résultats découlent de l'action posée suite à une prise de décision en fonction de ces nouvelles connaissances. On pourrait ainsi réduire le coût de production et augmenter les bénéfices et ce, en respectant les missions de l'entreprise.

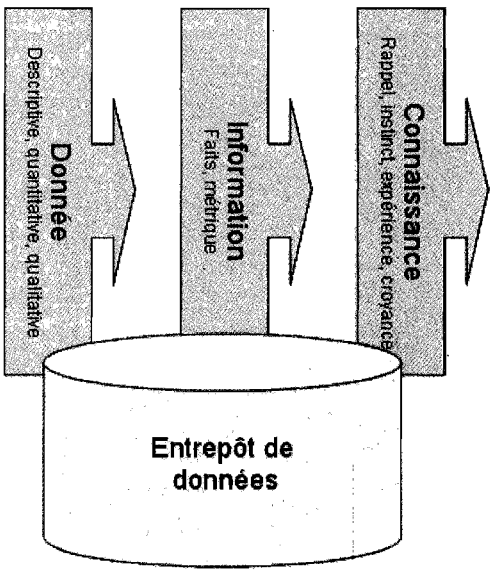


Figure 3.3 Étendue de l'entrepôt.

L'entrepôt de données s'étendra de la couche des «données» à la couche des connaissances (figure 3.3) afin d'aider les preneurs de décision à décider rapidement en leur permettant d'accéder facilement à l'information par des outils adaptés au BI. Certains outils vont même intégrer la couche «connaissance» et couvrir une partie de la couche «action».

À partir des sources de données des systèmes opérationnels, les chargements de l'entrepôt par les outils ETC permettront de faciliter la découverte de connaissances.

Selon [GARDNER 98], «un entrepôt de données est un processus, non un produit, qui assemble et gère les données de différentes sources afin de «gagner» une vue d'ensemble détaillée ou en partie de toute l'entreprise».

La figure 3.4 expose le cheminement des données de l'entrepôt. L'entrepôt contiendra des données prises dans les systèmes opérationnels (OPTL). Ces données seront intégrées, distribuées et livrées par un accès sécurisé aux différents types d'utilisateurs en passant directement par des outils de requête et d'analyse.

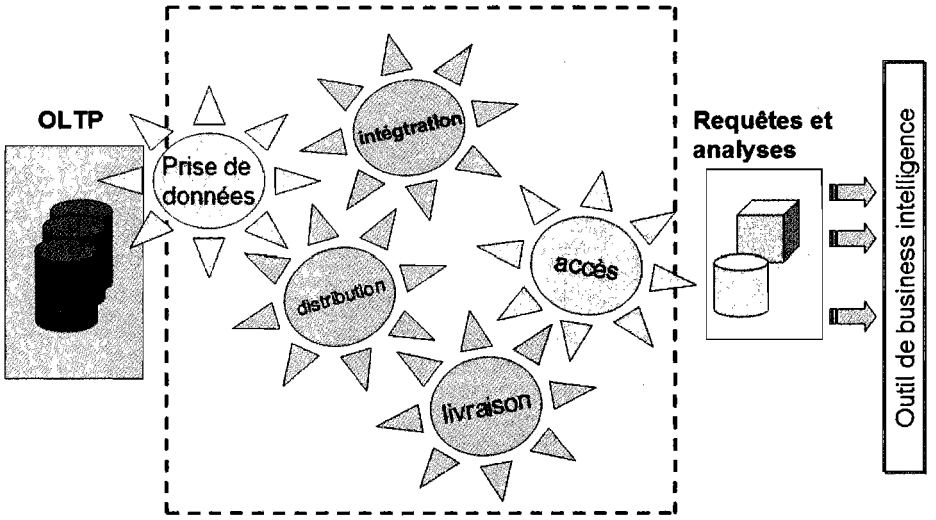


Figure 3.4 Schématisation des services de l'entrepôt.
Source : [TDWI 04]

En fusionnant les deux schémas, on peut voir l'emplacement de l'entrepôt et de ses services (figure 3.5). Les rôles de l'entrepôt s'étendront de la couche des «données» à la couche «action». Les outils de BI (*Business intelligence*) permettent donc à des personnes de prendre des décisions suite à l'analyse des résultats.

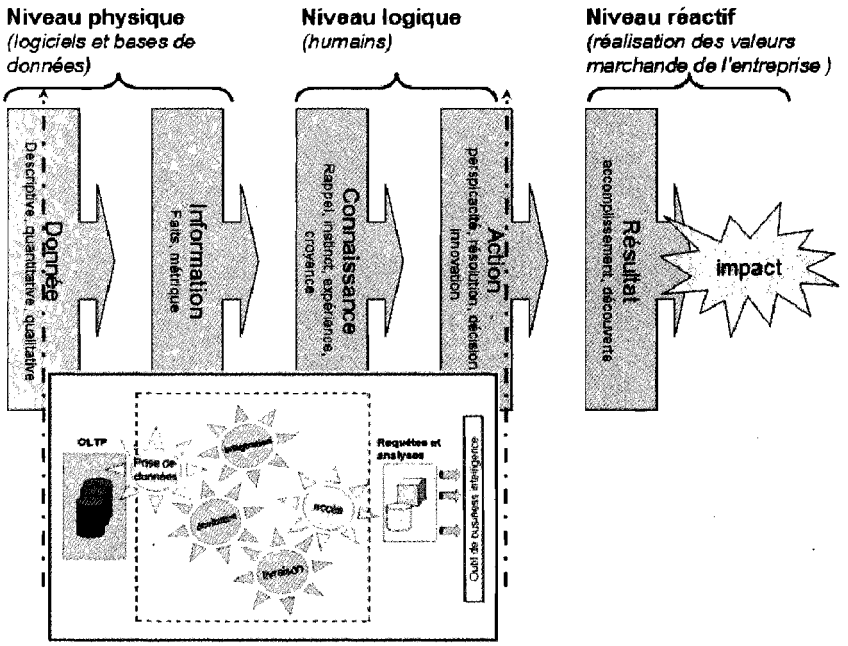


Figure 3.5 Positionnement de l'entrepôt et des outils de BI.

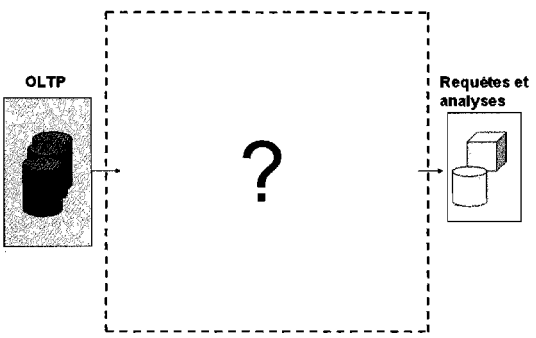


Figure 3.6 Vue schématique de l'entrepôt.

L'entrepôt (figure 3.6) se situe entre les données OLTP et les données de sortie pour les requêtes et les analyses. Il est représenté par le carré pointillé. L'entrepôt offre des services entre les sources de données qui l'alimentent et l'extraction de données qui seront transformées et présentées en sortie pour les requêtes et les analyses des demandeurs.

Pour réaliser cette intégration générale des processus de l'entreprise, des rôles sont définis. Chaque type d'architecture assignera ces rôles de façons différentes. Quels sont les rôles ou services d'un entrepôt de données. Selon [GARDNER 98] et [SHAHZAD 00], il existe cinq rôles essentiels. Voici dans le tableau 3.3 ces rôles et leurs définitions.

Tableau 3.3
Les rôles de l'entrepôt de données

Rôles	Description
Prise de données (Quête d'information)	Choix des données de différentes sources à extraire pour alimenter l'entrepôt. Choix des données exactes, nettoyées et complètes.
Intégration (Préparation des données)	(Rôle le plus important) Comment lier les données entres-elles ? Sont-elles intègres? Il faut effectuer un nettoyage, un calcul, une transformation ***et ce à tous les niveaux :*** clé, attribut, définition, structure, granularité, ...
Distribution (Chargement des données)	Publiciser l'entrée de nouvelles données dans l'entrepôt. (reliée au chargement réussi de l'ETC).
Livraison	Rendre accessible les bonnes données aux bons utilisateurs.
Accès (Interface sécurisé)	Accessibilité des données, interface utilisateur convivial.

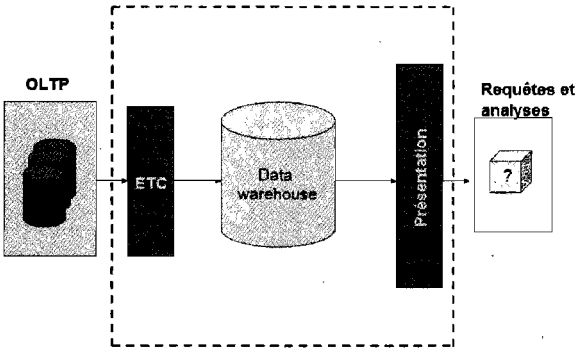


Figure 3.7 Service ETC et présentation.

Le premier rôle de l'entrepôt est d'effectuer la prise de données des systèmes OLTP. Choisir les données en fonction des besoins d'information requis.

Le second rôle est d'intégrer les données. Ce rôle sera attribué aux outils de chargement frontaux ou aux outils d'alimentation qui permettront l'extraction, la transformation et le chargement des données dans l'entrepôt. Il est représenté dans la figure 3.7 par le service ETC (ou ETL).

Tout le travail et le temps de développement reposent sur les outils ETC (50 à 70% du temps de développement). Le service ETC est réparti en trois phases. La phase d'identification et d'épuration des données consiste à définir et à identifier la donnée la plus pertinente en fonction de sa source. La phase de transformation regroupe les opérations de mise en format des données, de calculs des données secondaires et de fusion ou d'éclatement des informations composites. Enfin, la phase de chargement a pour rôle de stocker les informations de manière correcte dans les tables de faits de l'entrepôt de données.

Les deux rôles suivants, la «distribution» et la «livraison», veillent au bon déroulement de l'alimentation vers la présentation.

Enfin, le dernier rôle «Accès» est de permettre l'accès aux données. Le service de présentation des données est le seul point d'accès à l'entrepôt par les utilisateurs externes. Aucune demande de données ne pourra se faire par aucun autre service de l'entrepôt. On peut comparer cette phase à la cuisine d'un restaurant. Les clients ne vont jamais commander leur plat directement aux cuisiniers, il passe par le serveur qui lui en fait la demande aux cuisiniers. On parlera [TDWI 04] d'outils d'extraction du point de vue utilisateurs ce qui signifie l'extraction des données de l'entrepôt. À ne pas confondre avec l'étape d'extraction du service ETC qui lui alimente l'entrepôt.

3.4 Méthodologies de conception

Les systèmes OLTP sont «orientés-transaction» tandis que l'entrepôt de données est «orienté-sujet». Puisqu'ils sont différents, la méthodologie de modélisation de l'entrepôt se doit d'être différente. Que choisir entre entrepôt de données et magasin de données.

À la figure 3.8, l'article de Vassiliadis [VASSILIADIS et al. 01] propose une modélisation à trois niveaux : la perspective conceptuelle, la perspective logique et la perspective physique.

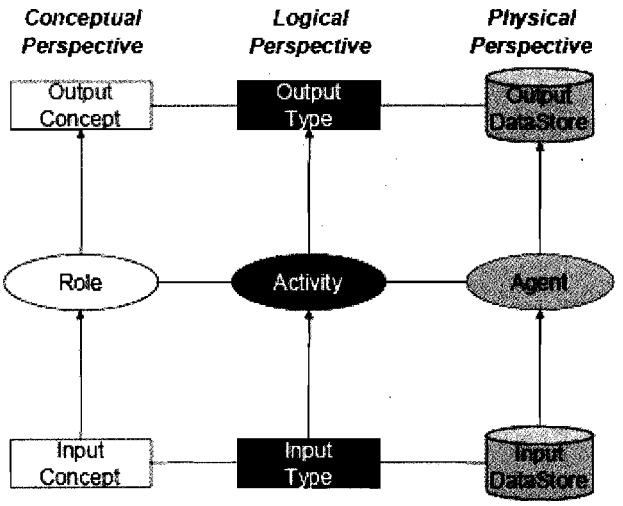


Figure 3.8 Le raisonnement en trois perspectives pour la modélisation.
Source : [VASSILIADIS et al. 01]

Un autre article [SHUNUNG et al. 05] propose plutôt une méthode de modélisation à quatre niveaux pour le développement d'un entrepôt de données. Puisque la méthode de modélisation à trois niveaux ne tient pas compte de l'orienté-sujet ni de la description des métadonnées dans les entrepôts, cette nouvelle méthode renforce et améliore la clarté et la modularité de l'entrepôt. Cette méthode est de type «*top-down*», «orientée-sujet» et se base sur les besoins des utilisateurs.

Les auteurs [SHUNUNG et al. 05] comparent les méthodes de modélisation des entrepôts de données et des bases de données. Le tableau 3.4 compare les étapes des méthodologies à trois et à quatre niveaux.

Tableau 3.4
Comparaison des étapes de modélisation

Nom des niveaux	Entrepôt de données	Base de données
Modèle conceptuel	Diagramme des processus *(nom, dimension, niveau, type et mesure)*	Diagramme de flux
Modèle logique	Schéma étoile	Schéma ER
Modèle objet	Arbre du sujet	----------
Modèle physique	Structure de données physique	Structure de données physique

L'étape ajoutée pour la modélisation de l'entrepôt de données se trouve au niveau du modèle objet : l'arbre du sujet que l'on peut voir à la figure 3.9a . En conceptualisant le processus d'affaires, on découpe le sujet en faits, en mesures et en dimensions. Par la suite, si elle existe, on définit la hiérarchisation de la dimension. Le modèle physique reposera sur le modèle objet qui est en fait un découpage en modèle dimensionnel.

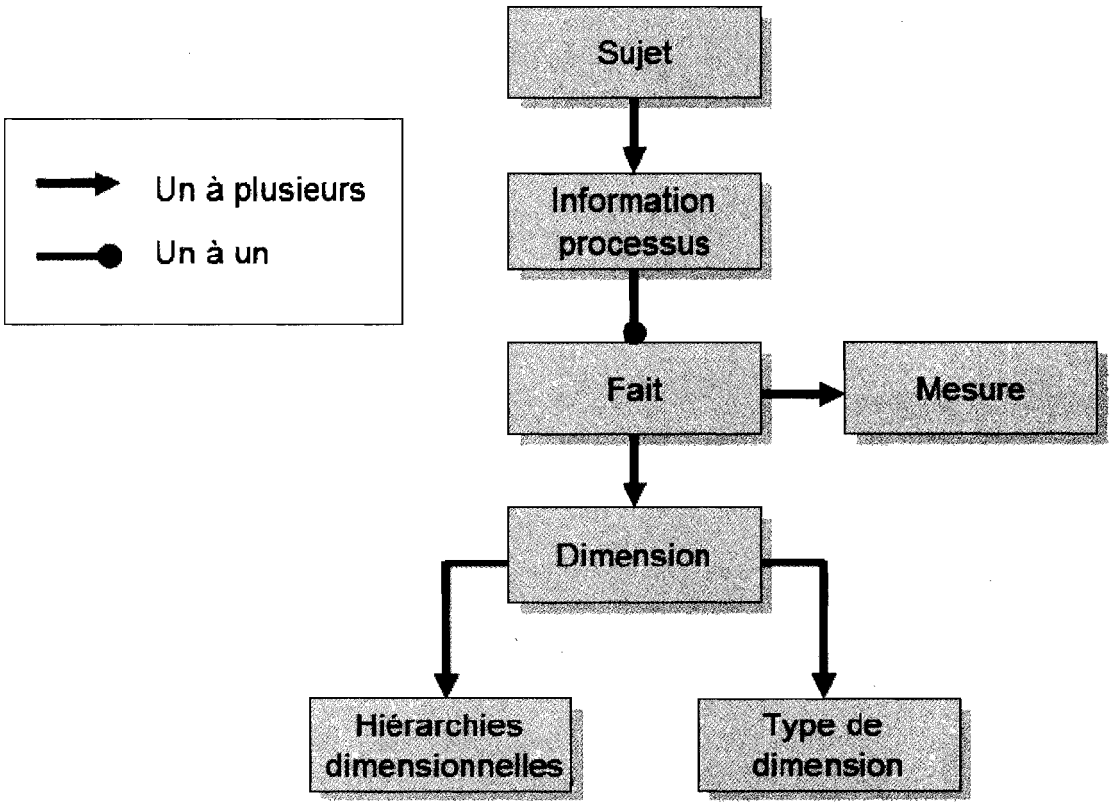


Figure 3.9a Schématisation de l'arbre du sujet.
Source : [SHUNUNG et al. 05]

Dans le même ordre d'idée [GOLFARELLI-RIZZI 98], Golfarelli et Rizzi ont défini une nouvelle méthode de schématisation du modèle dimensionnel. La figure 3.9a expose la méthode DFM (*Dimensional Fact Model*). Cette méthode est centralisée sur la table de faits. En se basant sur cette table, chaque dimension est décomposée en lignes et en ronds. Les ronds représentent la hiérarchisation de la dimension. On placera sous le point le nom de la hiérarchie de niveau inférieur suivi des autres niveaux (semaine, mois, année). La ligne pointillée détermine les opérateurs de fonction possibles pour la mesure de la table de faits (minimum et moyenne). Lorsqu'une ligne simple origine d'un rond sans être reliée à un autre rond, c'est qu'on peut représenter au même niveau hiérarchique d'autres informations sur la dimension (produit : poids ou nombre d'unités par boîte ou la grosseur du paquet).

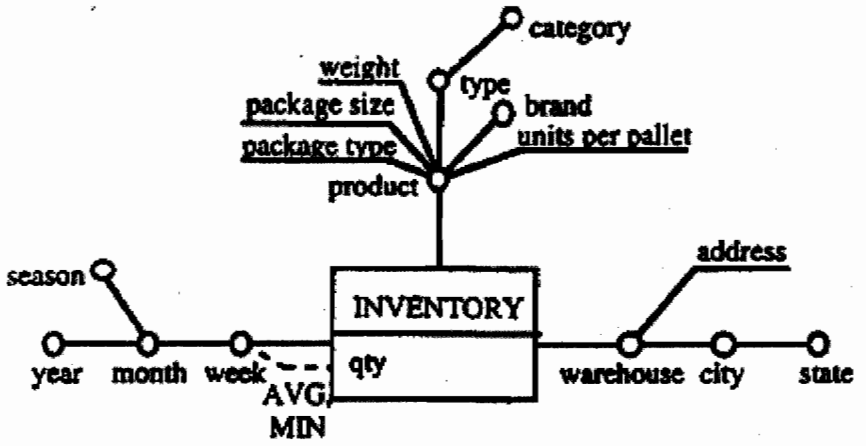


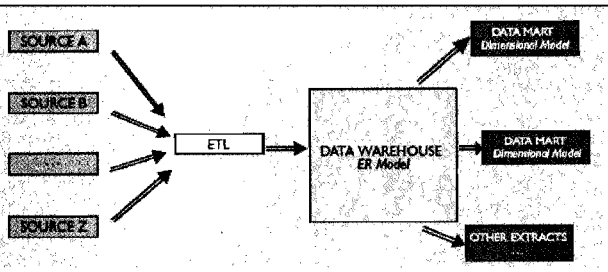
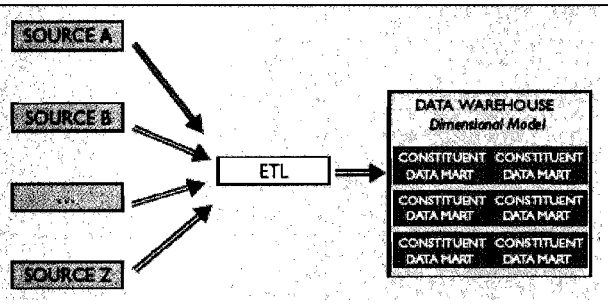
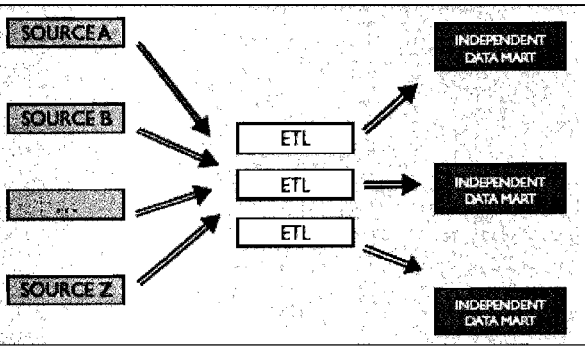
Figure 3.9b Schématisation du modèle dimensionnel avec DFM.
Source : [GOLFARELLI-RIZZI 98]

Chaque table de faits peut être représentée de cette façon. La superposition de faits est aussi possible si la hiérarchie des dimensions est la même.

Un autre article de synthèse sur la modélisation des entrepôts de données [LENZ et al. 03] recensait une méthode basée sur UML qui était aussi utilisée pour la modélisation de tout l'entrepôt.

Selon [JURIC 06] et [GARDNER 98], il existe trois options de modélisation logique représentées par les schémas suivants :

Tableau 3.5
Architecture logique de l'entrepôt

	**Modélisation E-R (Inmon)** L'entrepôt est modélisé en modèle «entité-relation» comme dans les bases de données opérationnelles afin d'avoir une centralisation des données. Des *data marts* sont créés par la suite.
	**Modélisation dimensionnelle (Kimball)** Dénormalisation et modification de la structure transactionnelle en plusieurs *data marts* **liés** entre eux.
	**Data marts indépendants** Modification de la structure transactionnelle en plusieurs *data marts* qui ***ne sont pas liés*** entre eux souvent nommés «tuyau de poêle». Donc, ici, il n'y a pas de véritable entrepôt.

Légende			
Système OLTP	**ETL (ETC)**	**Data Warehouse**	**Data mart**
Source des données	Extraction, transformation et chargement	Endroit où sont sauvegardées les données	Sous-division en sujet de l'entrepôt
SOURCE A	ETL	**Structure des données de l'entrepôt**	DATA MART

3.4.1 Explication de la modélisation ER (Inmon)

L'entrepôt est modélisé en «entité-relation» (ER) comme dans les bases de données transactionnelles afin d'avoir une centralisation des données. Il est donc piloté par la cible, ce qui veut dire par les données. La figure 3.10 schématise le type d'architecture ER d'Inmon. Les magasins de données ou comptoir d'information (*data marts)* sont créés par la suite, une fois l'entrepôt construit.

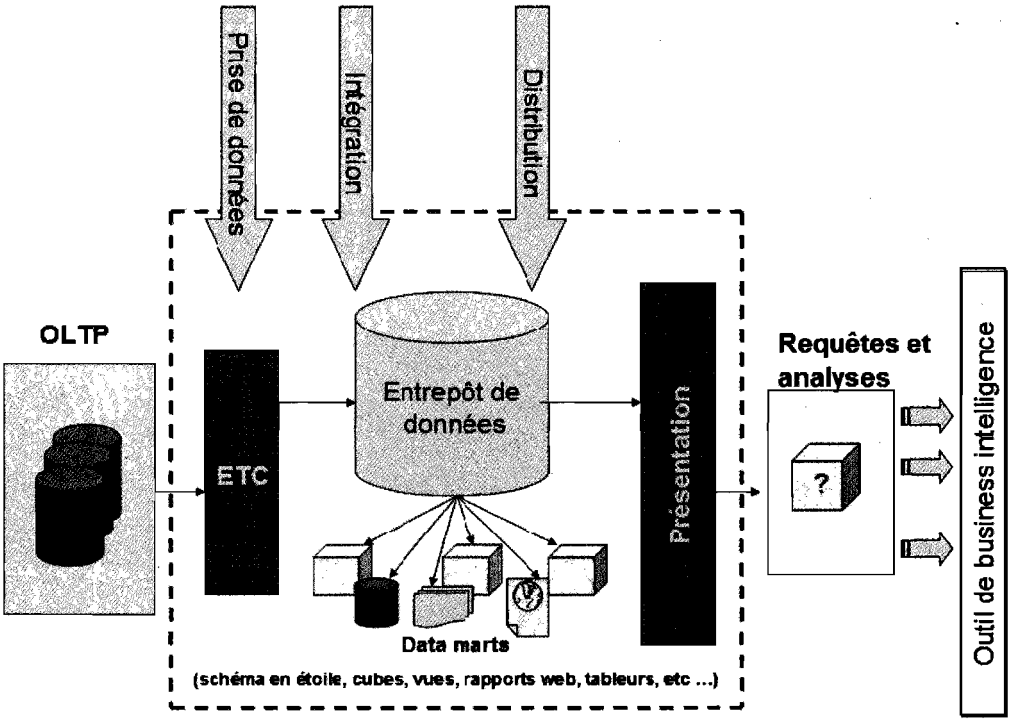


Figure 3.10 Entrepôt de données de type «Inmon».

Après qu'un tel entrepôt de données soit créé, il sert alors de source de données pour les cibles telles que les marchés de données dimensionnelles et pour toutes autres cibles non dimensionnelles (Ex. : tableurs, rapports web). Ces cibles sont alimentées par le modèle ER de l'entrepôt et porte l'appellation de «tuyau de poêle». Dans cette approche, les services (rôles) offerts sont la prise de données, l'intégration et la distribution.

Avantage :

- La méthode d'Inmon est plus puissante si d'autres types de magasins analytiques de données étaient nécessaires rapidement : un schéma ER, extraits de tableur, de *data set* pour le *data mining*, fichiers plats etc.

Inconvénient :

- Construire tout l'entrepôt avant de débuter le premier *data mart*.

3.4.2 Explication de la modélisation dimensionnelle (Kimball)

Le schéma ER de la structure transactionnelle est dénormalisé en plusieurs magasins de données liés entre eux. On parle alors de faits et de dimensions conformes. La figure 3.11 schématise le type d'architecture dimensionnel de Kimball.

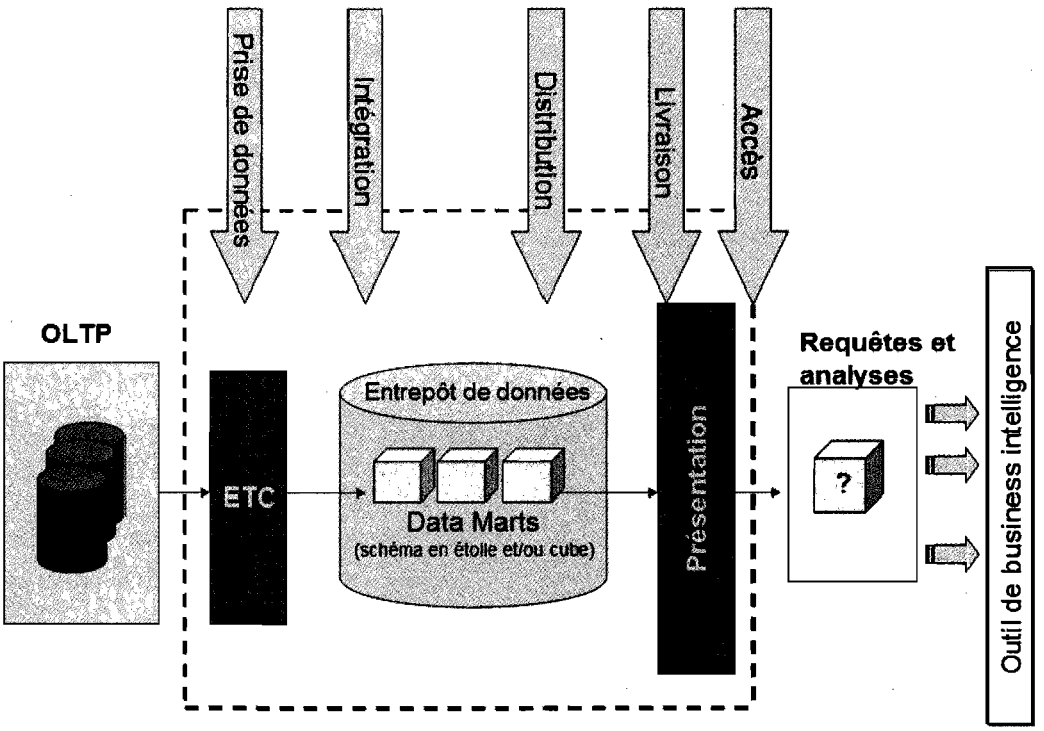


Figure 3.11 Entrepôt de données de type «Kimball».

Cette approche est analogue à l'approche précédente quand elle vient à l'utilisation des points d'émission de données opérationnelles et du processus d'ETC. La différence est la technique utilisée pour modéliser l'entrepôt de données. Dans cette approche, un ensemble de dimensions généralement utilisées (tel que le calendrier) connues sous le nom de dimensions conformes est d'abord conçu. Des tables de faits correspondant au sujet de l'analyse sont alors ajoutées. Un ensemble de modèles dimensionnels est créé où chaque table de faits est reliée aux dimensions multiples, et certaines des dimensions sont partagées par plus d'une table de faits.

Le résultat est un entrepôt de données qui est une collection de marchés dimensionnellement modelés, entrelacés de données conformes. Cette approche porte l'appellation de «bus de données».

Avantages :

- La rapidité de création car on peut ajouter un *data mart* à la fois.
- La simplicité de compréhension par les utilisateurs car la dénormalisation permet de mieux correspondre à leur façon intuitive de penser au lieu de présenter une structure normalisée où la structure des tables peut être complexe.

Inconvénients :

- Intégration de nouveau *data mart* sur un sujet ayant des points communs avec un autre *data mart* mais aussi, par exemple : différentes définitions pour un même domaine. Il faut donc statuer sur une définition unique.

3.4.3 Explication de la modélisation des magasins de données indépendants

Dans ce type d'architecture, le schéma ER de la structure est transformée en plusieurs *data marts* qui ne sont pas liés entre eux (souvent nommés «Tuyaux de poêle»). Il n'y a aucun auteur connu qui revendique le titre de ce type d'approche.

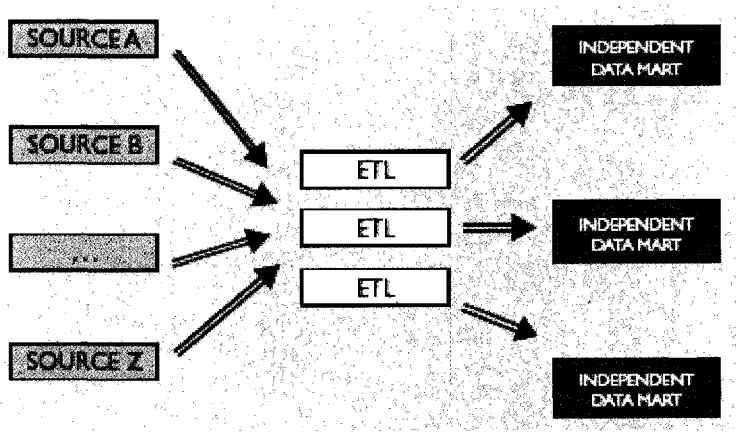


Figure 3.12 Entrepôt de données avec *data marts* indépendants.

Puisque les *data marts* sont indépendants, les mêmes données peuvent être présentes dans plusieurs *data marts*. Cette répétition de données duplique l'effort ETC inutilement. De plus, il en résulte d'une incapacité pour l'analyse à croiser les différents *data marts*.

Avantages :

- Fortement orienté-sujet.
- Personnalisation des analyses pour les différents utilisateurs.

Inconvénients :

- Répétition d'informations et de traitements.
- Aucune analyse croisée fiable.

3.4.4 Cycle de vie

Le cycle de vie d'un logiciel représente toutes les étapes de son développement et de sa maintenance. Il s'assure que chaque étape est réussie avant de passer à la suivante. Le but de ce découpage par étapes est d'en contrôler les erreurs afin de minimiser les coûts.

Pour l'entrepôt, un cycle de vie décisionnel existe aussi pour les mêmes raisons. Le cycle de développement d'un entrepôt de données nommé X-Meta par [CARNEIRO-BRAYNER 00] est représenté par la figure 3.14. Il se divise en trois grands axes : l'introduction, le développement et la production.

Lors de la création d'un prototype, seule la phase d'introduction est réalisée. La figure 3.13 montre en détail les phases d'introduction d'un prototype.

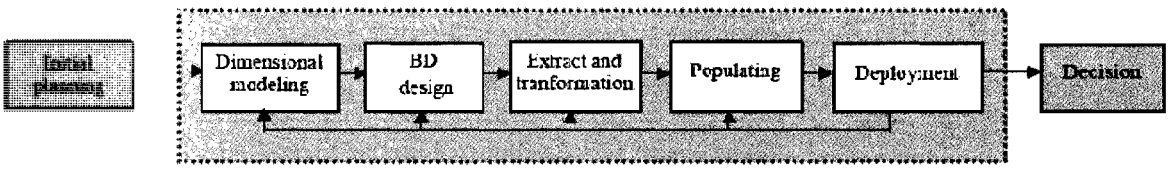


Figure 3.13 X-Meta : Phases composant l'introduction d'un prototype.
Source : [CARNEIRO-BRAYNER 00]

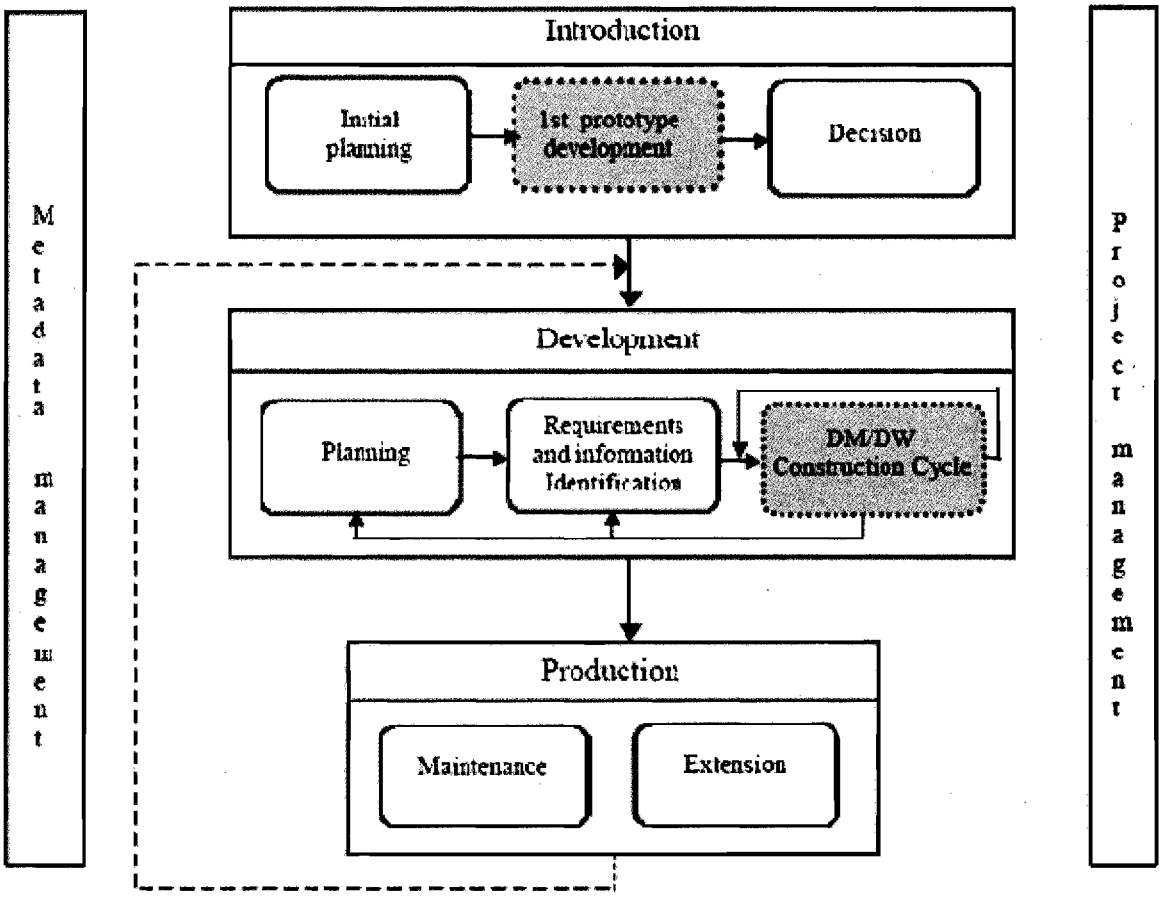


Figure 3.14 X-Meta : Cycle de vie décisionnel.
Source : [CARNEIRO-BRAYNER 00]

La phase la plus imposante de la méthode X-Meta se retrouve au niveau de l'axe «développement». Le module « DM/DW Construction Cycle» présenté à la figure 3.15 de la page suivante est divisé en petites parties réparties en quatre groupes parallèles. Tous les groupes sont alimentés par la phase précédente soit «*Requirements and Information Identification*». Dans le processus de développement, on pourra parcourir l'axe d'un groupe, de deux groupes ou de tous les groupes. Ce processus est itératif.

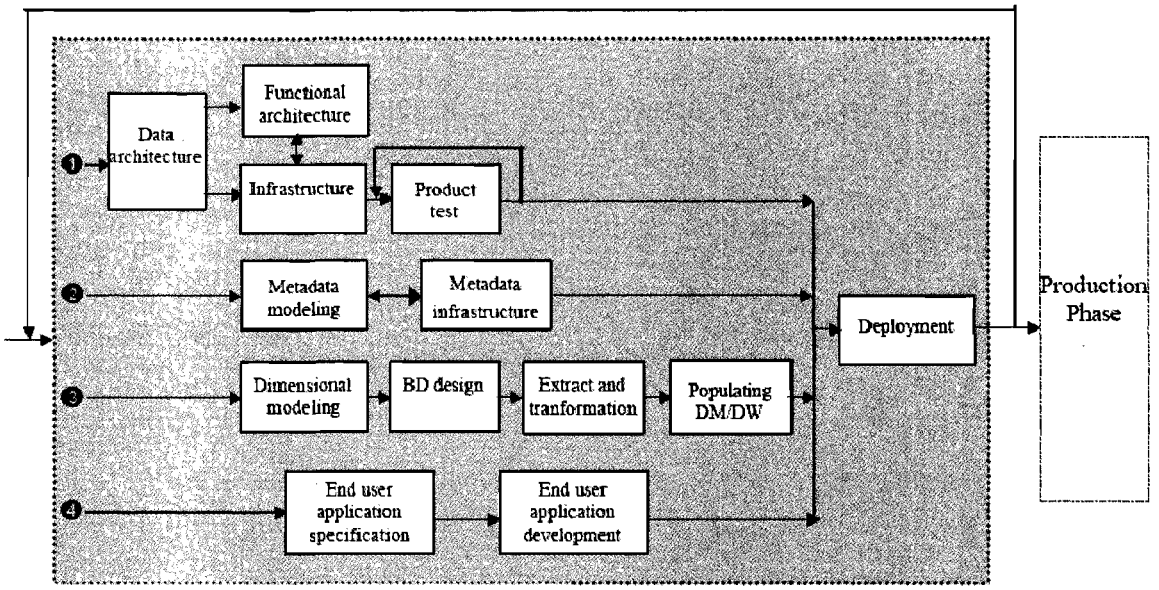


Figure 3.15 Détail X-Meta de la phase du développement du cycle des *data marts.*
Source : [CARNEIRO-BRAYNER 00]

Comparons avec la méthode du cycle de vie de Kimball. La figure 3.16 démontre les diverses étapes du « cycle de vie » pour la réalisation d'un entrepôt de données selon Kimball [Kimball *et al.* 05].

Brièvement, excluant les phases de déploiement et d'entretien, le processus comprend trois axes principaux qui peuvent être parcourus de façon indépendante :

- l'axe technique : architecture technique et sélection des produits;
- l'axe des données : modélisation dimensionnelle et zone temporaire de traitement;
- l'axe d'analyse : spécification et réalisation d'applications d'analyse.

Comme le montre la figure 3.16, l'étape de la «définition des besoins d'affaires» est un pré-requis à tous les axes du modèle de cycle de vie.

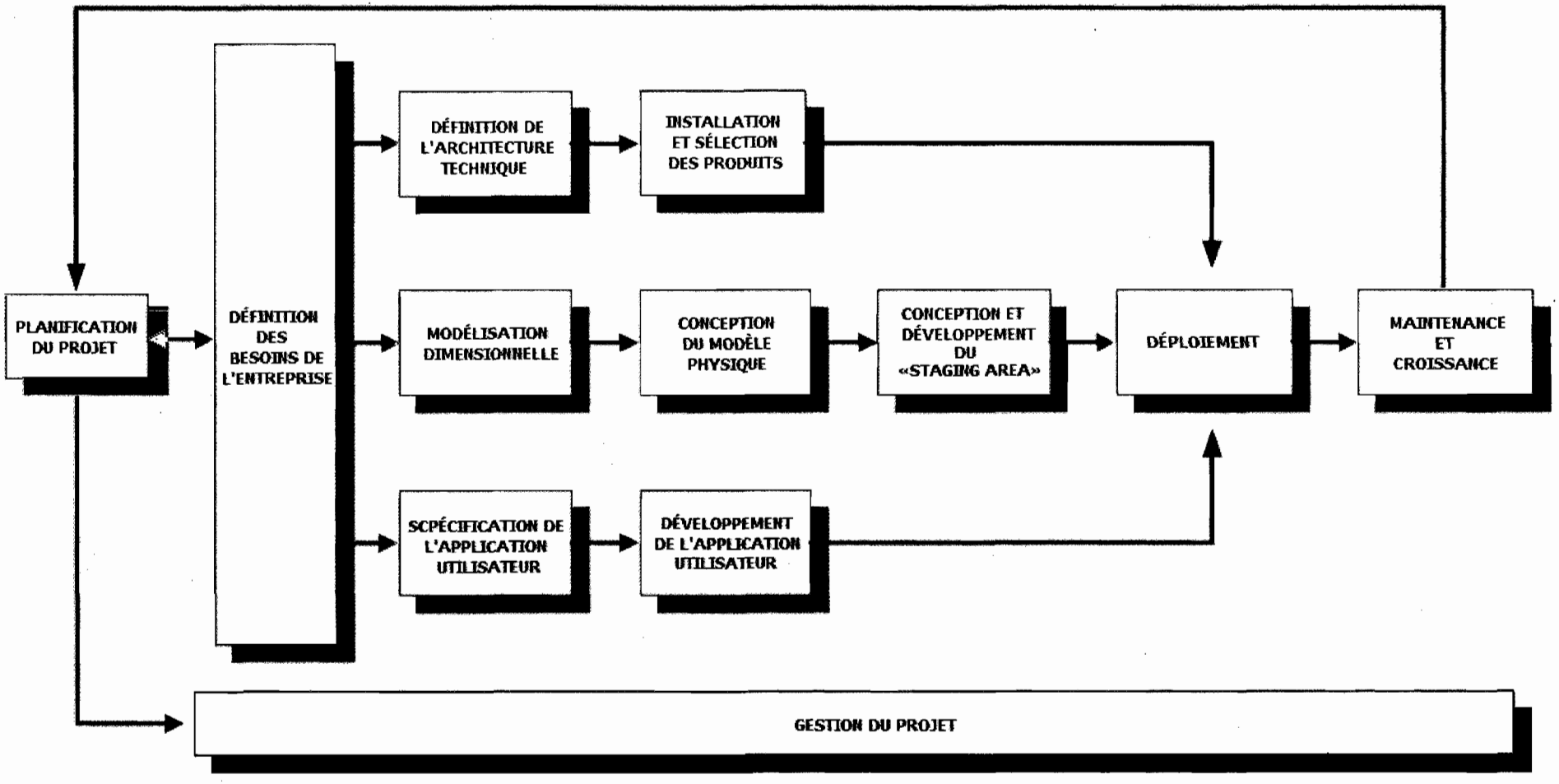


Figure 3.16 Cycle de vie dimensionnel de Kimball.
Source :[Kimball et al. 05]

3.4.5 Phases communes du développement

Il existe des phases communes [SEN-SINHA 05], [CARNEIRO-BRAYNER 00], [Kimball et al. 05], [TDWI 04] entre les différentes méthodes de cycle de vie de réalisation d'un entrepôt. Le tableau 3.6 nomme et explique les sous-tâches associées à chaque phase.

Tableau 3.6
Phase de développement de l'entrepôt (Source : [TDWI 04])

Phases	Sous tâches à réaliser
Analyser les besoins de l'entreprise	➢ Sondage et questionnaire (brainstorming) afin d'identifier les besoins à des fins de prise de décision et d'analyse. ➢ Prioriser les besoins utilisateurs. ➢ Estimation des risques associés à la prise de décision d'un besoin.
Le modèle conceptuel de données	➢ Création de la solution de chaque besoin pour obtenir le schéma de l'organisation. ➢ «Data design» qui comprend le modèle des données ainsi que la normalisation. ➢ La modélisation de l'entrepôt.
Le modèle d'architecture	➢ L'architecture est le schéma permettant les communications, la planification, la maintenance, l'apprentissage et la réutilisation. ➢ Il inclut différents espaces : le modèle de données, le modèle technique et le logiciel/matériel. La façon de les définir peut être *top-down, bottom-up, hybride* ...
L'implémentation	➢ Sélection des sources de données. ➢ Application des transformations pour charger l'entrepôt. ➢ Permettre par une interface de présentation aux usagers une prise de décision et/ou une extraction. ➢ Il est important de s'assurer de la qualité des données de bout en bout. ➢ Il est aussi important de bien documenter les données : la gestion des métadonnées (définitions, faits, nettoyage, ETC).
Le déploiement	➢ Faire les ajustements, les tests sur un premier besoin. ➢ Prévoir la maintenance de cette première version. ➢ Alimenter / publier les données. ➢ Il y aura plusieurs versions de l'entrepôt. ➢ Prévoir l'évolution des besoins de l'entrepôt.

3.5 Les métadonnées

Les métadonnées représentent le dictionnaire unique de l'entrepôt. Ce répertoire conserve les descriptions de l'ensemble des fonctionnalités et uniformise la définition des concepts et des attributs. Les métadonnées [SHAHZAD 00] décrivent aussi les processus et les règles de transformation des données sources des OLTP vers les données cibles de l'entrepôt. Voici ce que définissent les métadonnées :

- Les éléments de données et leurs types.
- La définition d'affaires des éléments de données.
- Comment et quand mettre à jour les données.
- La définition des éléments ayant le même sens.
- Les valeurs valides de chaque élément de données.
- Les règles de transformation (source vs cible).

Comme exemple concret de métadonnées, on retrouve les listes des transformations et les listes des chargements des données. L'information sur la construction de règles d'entreprise appliquées à l'entrepôt est aussi considérée. En tant que dictionnaire de données, toutes les données et les structures y sont répertoriées avec une carte de localisation des informations (source versus cible). Chaque donnée et chaque structure ont leur définition technique et leur définition conceptuelle pour l'utilisateur. On y définit aussi les règles de sécurité et d'accès aux données.

L'aspect temporel des modifications des structures fait aussi partie des métadonnées. Il sera abordé au point suivant.

La plupart des outils ETC ne conservent que les définitions des tables et des champs. Il faut jongler avec les fonctionnalités pour obtenir directement les listes des transformations et des chargements, le lien entre les cibles et les sources. Cette question sera abordée plus profondément au chapitre 5.

Les métadonnées représentent l'élément capital de nos préoccupations. Elles sont essentielles à la construction de l'entrepôt, à son entretien, à son suivi, pour ne pas dire à

sa survie, et à son extensibilité. Elles représentent la documentation technique requise lors du développement logiciel.

3.6 L'aspect temporel de l'entrepôt

Lorsque l'on parle d'entrepôt de données temporelles, on parle d'historique des données. Que veut dire le terme «données historiques». Que chaque nouvelle insertion de données provenant des systèmes transactionnels n'efface pas les anciennes valeurs, mais crée une nouvelle occurrence de la donnée.

L'aspect temporel permet de répondre facilement à ce genre de question :

- Quand un étudiant a-t-il fait sa première demande ?
- Combien d'étudiants ont fait une demande d'admission depuis le 1er mars 2000 ?
- Quel a été le nombre maximum d'étudiants admis dans le programme XXXX entre 2000 et 2005 ?
- Quel fut le programme le plus en demande pour les étudiants provenant de la Mauricie ?

Il y a trois points à résoudre pour la dimension temporelle :

- Comment représenter les données historiques ?
- Comment gérer des données historiques ?
- Comment interroger des données historiques ?

Observons d'abord le premier point «comment représenter les données historiques?». L'échelle de temps logique est représentée par une valeur discrète numérique partant de 0. La figure 3.17 nous illustre les valeurs possibles de l'échelle de temps. Chaque unité représente la plus petite granularité nécessaire du temps. La valeur 1 est le 1er jour de vie de l'entrepôt de données. Le «NOW» est la valeur de la granularité actuelle. La valeur 9999 indique le futur c'est-à-dire que la valeur est toujours valide à ce jour. Il n'y a pas eu de changement sur cette donnée. Lorsqu'une nouvelle donnée est ajoutée, elle a la valeur

temporelle suivante : [NOW,9999[. La figure 3.17 représente la dimension temps au moment «NOW».

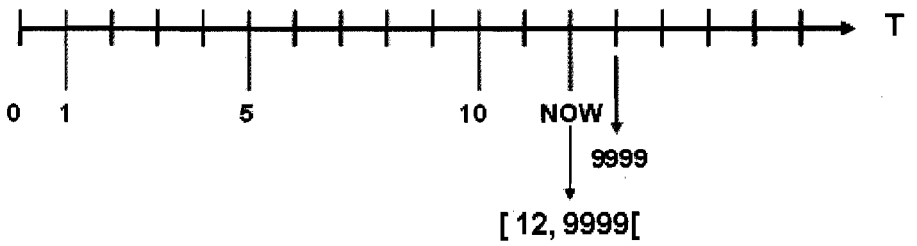


Figure 3.17 Représentation de la dimension temps.

Chaque donnée dans l'entrepôt est donc représentée par un couple temporel. La première valeur du couple est la variable nommée «H_DEBUT» qui représente le moment où la donnée a été introduite dans l'entrepôt. La deuxième valeur du couple est la variable nommée «H_FIN» qui nous indique que l'enregistrement n'est plus valide à compter de ce moment. La valeur 9999 représente le futur. Si «H_FIN» contient 9999, l'enregistrement est considéré comme toujours valide.

Si nous sommes au temps T=12, alors «NOW» = 12 est la façon de représenter cette information temporelle est **[**12,9999**[.**

Dans la structure, deux champs permettront de représenter les données historiques. Soient «DT_DEBUT» et «DT_FIN».

Observons maintenant «comment gérer des données historiques?». Le tableau 3.7 nous résume les modifications faites à un enregistrement modifié du côté des systèmes transactionnels versus les modifications temporelles dans l'entrepôt. Revenons sur le fait qu'un entrepôt de données temporelles ne fait qu'ajouter les nouvelles données. Lorsqu'une donnée est modifiée dans le système transactionnel, un ajout est fait dans l'entrepôt. Supposons un enregistrement E^1 intégré dans l'entrepôt au temps t^1, lors de son ajout, le couple temporel de cet enregistrement est représenté comme suit :

$$E^1 = [t^1,9999[$$

Si au temps t^2 l'enregistrement E^1 est modifié dans le système transactionnel on ajoutera à l'entrepôt le nouvel enregistrement $E^{1'}$. Il faudra alors informer la structure.

Tableau 3.7
Gestion des données temporelles sur un enregistrement modifié dans l'OLTP

Voici l'enregistrement E^1 avant l'opération	**$E^1 = [t^1, 9999[$**
Étape 1 : ajout au temps t^2 de $E^{1'}$	**$E^{1'} = [t^2, 9999[$**
Étape 2 : ajustement de la fin de validité de E^1 qui devient $E^{1'}$	**$E^1 = [t^1, t^2]$**
Voici les enregistrements après la gestion de cette nouvelle donnée	**$E^1 = [t1, t2]$** **$E^{1'} = [t2, 9999[$**

Il reste à déterminer «comment interroger des données historiques?». Lorsqu'une extraction de données est demandée par une requête, il faut connaître le moment précis ou la plage exacte de temps que l'on veut extraire pour pouvoir répondre avec les bonnes données.

Extraction au moment présent : Lorsque l'on veut extraire toutes les données valides lors du dernier chargement, le champ «H_FIN» est alors égal à 9999. Il suffit d'ajouter la clause suivante «where H_FIN = 9999» à la requête. On peut aussi avoir un champ «IND_NOW» qui contient 1 lorsque l'enregistrement est valide et 0 lorsque l'enregistrement ne l'est plus. Il suffit alors de modifier la clause somme suit : «where IND_NOW = 1».

Extraction à un moment ***t***: Lorsque l'on veut extraire toutes les données valides lors du chargement ***t***, il faut extraire toutes les données qui étaient présentes au temps ***t*** donc «H_BEDUT» <= ***t*** mais qui n'étaient pas supprimées au temps ***t*** donc «H_FIN» > ***t***. Il suffit d'ajouter la clause suivante à la requête : «where H_DEBUT <= ***t*** and H_FIN > ***t*** »

Extraction à l'intérieur d'une plage de données : Que faire si l'on veut extraire les données valides entre le moment t^1 et t^2. Dans ce cas de figure, la complexité augmente. Regardons les zones de validité possibles pour un enregistrement à l'aide de la figure 3.18.

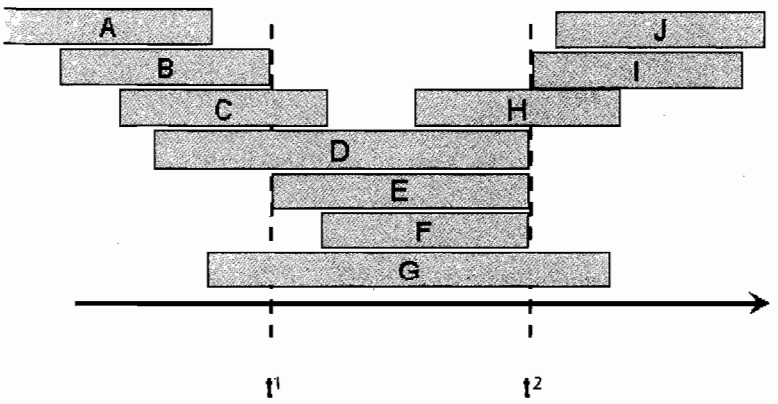


Figure 3.18 Zones de validité des enregistrements.

Pour déterminer les zones considérées pour l'extraction, il faut savoir ce que l'on veut faire comme opération avec les données. La table de faits déterminera lequelle des quatre cas de figure sera utilisée.

Il faut considérer que «H_DEBUT» est toujours plus petit que «H_FIN» et que t^1 est toujours plus petit que t^2.

«Extraire les données qui ont été valides durant toute la durée de la plage soit entre t^1 et t^2 *»*
Dans cette situation, seulement la zone G sera considérée puisque seules ces données sont valides pendant toute la durée de l'intervalle. La clause serait :

Zones : G,

where H_BEDUT <= t^1 AND H_FIN > t^2

«Extraire toutes les données qui ont été valides au moins une fois entre t^1 et t^2*»*
Dans ce cas de figure, seront excluses les zones A, B, I et J. La clause sera la suivante

Zones : C,D,E,F,G,H

Where (H_BEDUT < t^1 AND H_FIN < t^2 AND H_FIN > t^1) or
(H_BEDUT <= t^1 AND H_FIN >= t^2) or
(H_BEDUT <= t^1 AND H_FIN = t^2) or
(H_BEDUT > t^1 AND H_FIN = t^2) or
(H_BEDUT = t^1 AND H_FIN = t^2) or
(H_BEDUT > t^1 AND H_FIN > t^2 AND H_DEBUT <= t^2)

«Extraire toutes les données qui ont été valides au moins une fois entre t^1 et t^2 *excluant ceux dont la fin de validité est entre les deux moments»*

Zones : G,H

Where (H_BEDUT <= t^1 AND H_FIN >= t^2) or
(H_BEDUT > t^1 AND H_FIN > t^2 AND H_DEBUT =< t^2)

Les cas de figure pourront différer si l'on décide d'inclure ou d'exclure les bornes (t^1 et t^2). La plupart des logiciels OLAP offrent la possibilité d'une saine gestion de la dimension temps lors de la présentation des données.

Évidemment ce type d'événement temporel n'existe pas dans les OLTP. Il existe des processus qui permettent de représenter ces données non temporelles en données temporelles. Trois articles proposent différents points de vue pour la gestion de l'aspect temporel.

Selon Amo et Alves, [AMO-ALVES 00], un entrepôt de données temporelles peut être défini par un ensemble de vues matérialisées temporelles sur des sources de données non temporelles. Dans son article, il définit des opérateurs algébriques et des règles pour sa méthode. Ils proposent une méthode pour gérer l'aspect temporel des données qui consiste en trois points :

a) L'historique est entièrement emmagasiné à l'extérieur de l'entrepôt de données dans les vues matérialisées.
b) Des vues auxiliaires appelées «compléments» gardent les informations de changement.
c) Les changements temporels sont exécutés en évaluant seulement les relations auxiliaires et les compléments par des opérateurs algébriques relationnels.

Selon [SERNA-ADIDA 05] il est important dans le domaine médical d'obtenir l'information courante et l'information concernant l'historique. Dans les bases de données traditionnelles, une perte de données peut être fatale. Il peut être crucial de ne pas avoir accès à l'historique des données d'un patient. Le système d'aide à la décision logistique et

médicale (ADELEM) permet de répondre à ces contraintes. Selon l'architecture du projet ADELEM, la gestion de l'aspect temporel est gérée à l'extérieur de l'entrepôt de données. Ils proposent un schéma temporel pour l'entrepôt de données représenté à la figure 3.19.

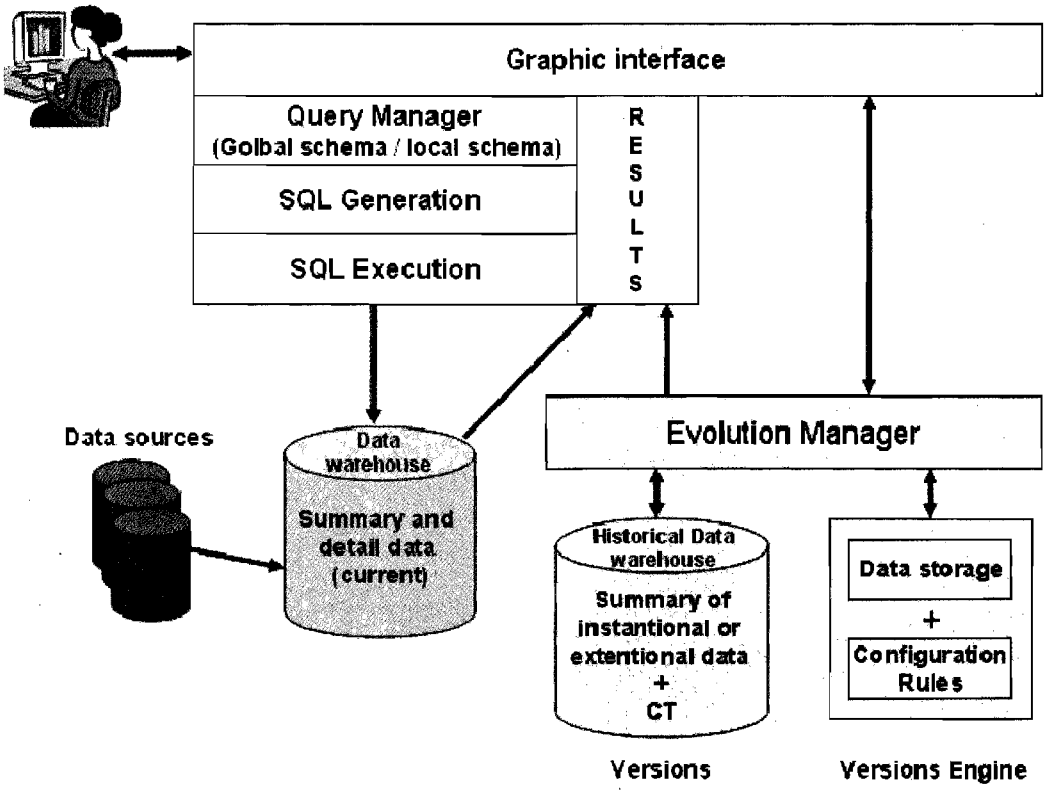


Figure 3.19 Architecture du projet ADELEM.
Source :[SERNA-ADIDA 05]

Voici les composants de son architecture :

- un opérateur appelé «SetVersion» permettant de suivre le changement de version;
- un ensemble de primitives pour la gestion des versions, des cubes et des dimensions;
- un gestionnaire d'évolution responsable de l'historique;
- un ensemble de règles définissant les opérations des primitives.

Dans un autre ordre d'idée, Eder [EDER-KONCILIA 02] propose une identification de la dernière version de la structure à la figure 3.20.

Pour chaque objet, une information permet d'identifier sa dernière période de validité. La version est définie comme suit : [ANNEE_DEBUT : ANNEE_FIN] où «ANNEE_DEBUT» représente l'année où la version de l'objet a débuté et ANNEE_FIN est la fin de validité de la version. Une dimension «structure» est créée permettant de suivre l'évolution des versions.

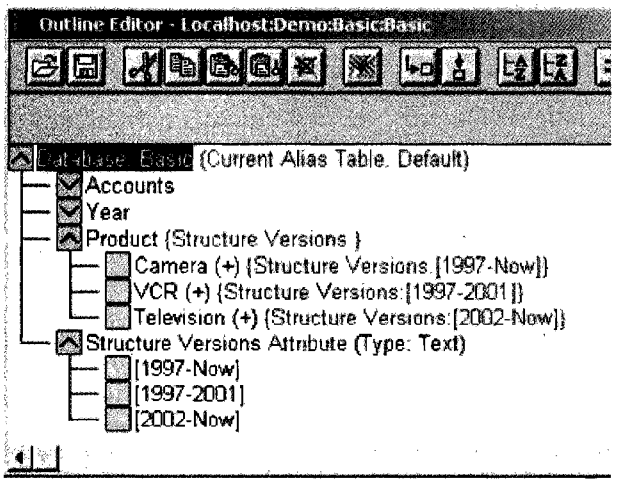


Figure 3.20 Version des structures.
Source : [EDER-KONCILIA 02]

Si une version est encore valide, elle n'a pas d'année de fin, on place la valeur «NOW» dans «ANNEE_FIN» (Ex. : [ANNEE_DEBUT : NOW]). Un objet peut être en relation uniquement avec une valeur de version qui doit être présente dans la dimension «structure». Eder, va ensuite étendre la notion de version aux attributs et aux enregistrements de l'entrepôt de données.

L'article de Wrembel [WREMBEL-MORZY 05] traite du même sujet. Il nomme cette approche MVDW (MultiVersion Data Warehouses) qui permet de contrôler les changements structuraux de l'entrepôt. C'est donc tout l'entrepôt qui est historisé. La version de l'entrepôt est composée d'une version de schéma et d'une version des instances. C'est par le MQL (Multiversion Query Languange) que les requêtes sont effectuées entre les différentes versions.

3.7 Modélisation dimensionnelle

L'étape de la modélisation des données n'est pas unique aux entrepôts de données. Selon [SHAHZAD 00], les raisons d'utiliser une modélisation sont les mêmes :

- définir l'étendue de l'entrepôt;
- permettre une vue d'ensemble de la complexité des relations entres les données;
- reconnaître et contrôler la redondance.

La piètre performance des systèmes transactionnels face aux requêtes complexes requises pour la prise de décision impose une autre modélisation pour l'entrepôt. Il existe trois types de représentation dimensionnelle [TRYONA et al. 99] : le schéma en étoile, le schéma en flocon et la constellation.

Un modèle dimensionnel est composé [JONES-SONG 05] de tables de faits et de table de dimension. Selon [Kimball et al. 005], la table de faits permet de mesurer l'activité et les tables de dimensions contiennent les informations faisant varier les mesures. Le modèle dimensionnel a été introduit par Ralph Kimball. C'est dans le modèle dimensionnel que seront emmagasinées les données en sortie des outils ETC.

Toujours selon Kimball, les structures dimensionnelles sont le fondement même de la mise en place des cubes OLAP. Elles sont simples à créer, stables et intuitivement compréhensibles par les utilisateurs finaux. Les quatre grandes étapes de la construction du modèle dimensionnel sont : «choisir le processus à modéliser », «définir la granularité du processus», «choisir les dimensions » et finalement «identifier les faits »

La figure 3.21 présente les trois (3) schémas dimensionnels existants : le schéma en étoile, en flocons et en constellation. Le plus courant est le schéma en étoile.

Le schéma étoile, selon [Kimball et al. 05], [SHAHZAD 00] et [JONES-SONG 05], contient une table centrale appelée «table de faits». Cette table est composée presque uniquement de clés primaires et étrangères. Les clés étrangères pointent vers les dimensions. Les dimensions sont placées tout autour de la table de faits.

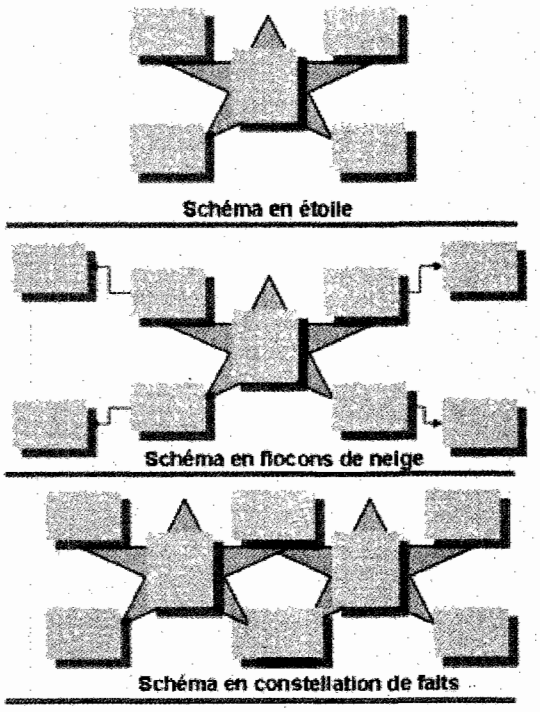


Figure 3.21 Représentation schématique des modèles dimensionnels.
Source : http://www.systemeETC.com/concepts_md.htm

Dans un autre ordre d'idée, les auteurs [TRYONA et al. 99] proposent un modèle «étoile – entité – relation» permettant la modélisation des dimensions en un schéma ER. Les symboles pour les connecteurs sont différents mais le principe est le même que l'ER soit des liens entre les dimensions et les faits.

Dans le livre de Kimball, [Kimball et al. 05] trois types de tables de faits sont répertoriés : table de faits avec mesure, table de faits sans mesure et table de faits avec événements.

3.8 Entrepôt, OLAP, DSS et *Data Mining*

L'entrepôt est un support d'aide à la décision. Dans un premier temps, les systèmes transactionnels accumulent de façon quotidienne des données opérationnelles. Par la suite, le gestionnaire peut analyser ces données. C'est l'entrepôt qui l'aidera à la prise de décision. Puisque la modélisation d'un entrepôt est plutôt à des fins d'analyse comparativement aux OLTP où la modélisation donne un accès rapide à un enregistrement, la performance est de mise. Pour obtenir cette performance, le modèle dimensionnel est utilisé.

Les systèmes d'analyse OLAP permettent l'agrégation, le forage et la coupe de données en dé (*dicing*) ou en tranche (*slicing*). Les cubes OLAP sont modélisés à l'aide du schéma dimensionnel. Un tableau énumérant les différents systèmes intégrant l'OLAP et leur moteur d'inférence (serveur) se trouve à l'annexe C. Les systèmes d'analyse OLAP reposent sur un serveur de type ROLAP, MOLAP, DOLAP, SOLAP ou HOLAP. La figure 3.22 nous montre quelques architectures serveurs.

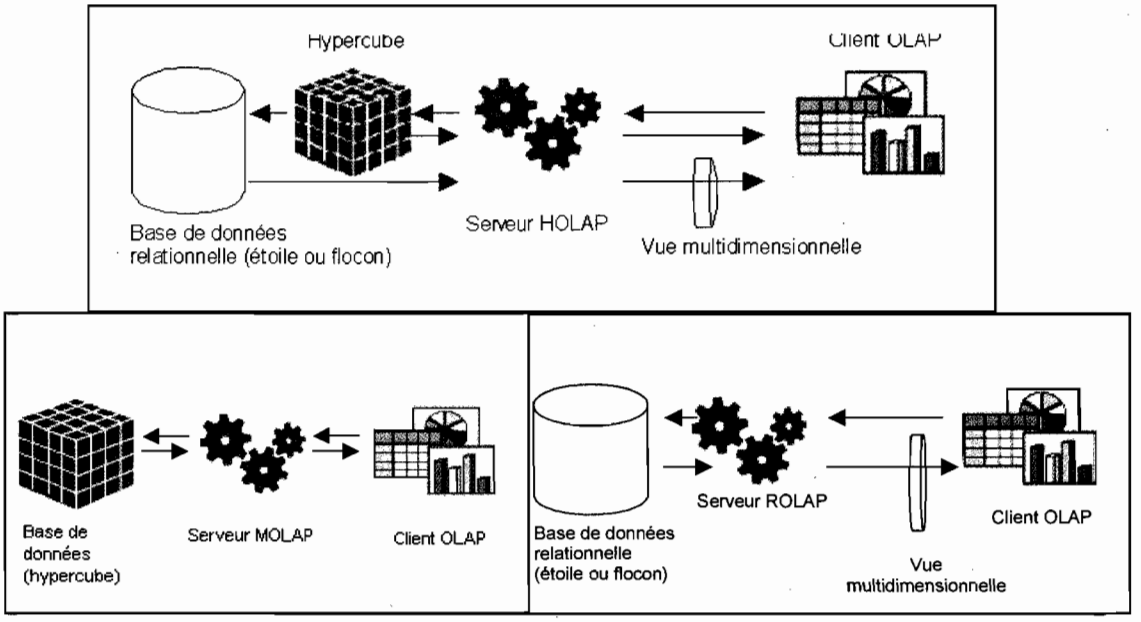


Figure 3.22 Type de serveur OLAP.
Source : «Introduction au projet SOLAP» (Université Laval)

Les données historiques emmagasinées dans l'entrepôt, [INMON 02] permettent aux entreprises de comparer sur plusieurs années l'évolution du marché. Elles peuvent analyser ces données historiques d'une autre façon. Cette analyse longitudinale permet, par exemple, d'expliquer les hausses et les baisses de la clientèle étudiante. En fouillant dans les données et en croisant divers critères de façon exploratoire, on arrivera à dégager des modèles de clientèles. Ces modèles permettront à leur tour de prédire les nouvelles clientèles. On parle alors ici de forage de données (*data mining*). Il se cache souvent une richesse insoupçonnée dans les données temporelles de l'entrepôt.

Des logiciels de forage de données permettent la fouille de données et la découverte de connaissances. Ces logiciels s'alimentent d'un fichier plat extrait de l'entrepôt. Si l'entrepôt est construit pour l'OLAP, il manquera uniquement un module de conversion du modèle dimensionnel en fichier plat afin de permettre à ces logiciels d'analyser les données en profondeur.

Trois logiciels de forage de données ont été explorés en 2006 : Oracle DataMiner (Oracle), Entreprise Miner (SAS) et Weka. Les deux premiers furent classés au stade de prototype et ne convenaient pas aux attentes et aux besoins de notre projet. Le dernier, Weka,

logiciel «Open source», a permis une évaluation satisfaisante de l'outil et pourra être utilisé pour de futures explorations.

3.9 Les outils

Des logiciels ou progiciels sont nécessaires à la conception d'un entrepôt de données. La finalité repose sur trois outils distincts : *des outils de base de données*, *des outils d'alimentation (ETC)* et *des outils de présentation*.

Tableau 3.8
Les bases de données les plus courantes

Oracle	www.oracle.com
IMB DB2	www.ibm.com
My SQL	www.mysql.com
Microsoft SQL Server	www.microsoft.com/sql
Sybase	www.sybase.com
NCR	www.ncr.com

L'article de [SEN-SINHA 05] en annexe D met en comparaison différentes bases de données. Le tableau 3.8 présente les bases de données les plus courantes.

La plupart des outils ETC d'aujourd'hui possèdent une interface graphique permettant le «glisser/déplacer» pour construire le flux de transformation et d'intégration des données vers l'entrepôt. Il existe des outils propriétaires et des outils «open source» sur le marché.

Les auteurs [FAN-POULOVASSILLIS 03] proposent une méthode nommée AUTOMED qui permet de décrire les flux de transformations des systèmes hétérogènes afin d'avoir une vue d'ensemble des données intégrées à l'entrepôt provenant de différentes sources.

Un autre auteur, [SIMITSIS 05], propose une méthode de conception des modèles logiques pour les processus ETC. Il définit ses symboles et modélise la transformation des données en étapes.

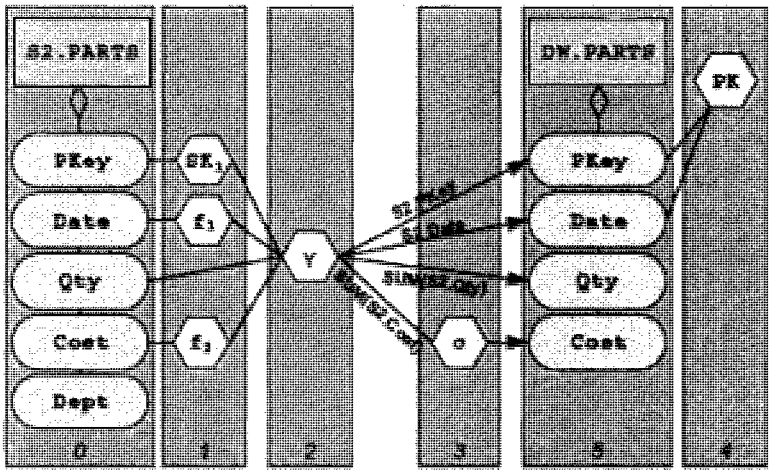


Figure 3.23 Modélisation du processus ETC.
Source : [SIMITSIS 05]

L'outil de présentation est la source qui permet à l'utilisateur de communiquer avec l'entrepôt de données. Il se doit d'être intuitif et simple d'utilisation. Un outil de présentation peut gérer différentes sorties. Il existe aussi des outils propriétaires et des outils «open source» sur le marché. Ils seront analysés au chapitre 5.

3.10 Travaux récents

Dans les travaux plus récents, on aborde déjà la notion d'entrepôt de données de prochaine génération. Un premier article de [PEIPERT-ALBALA 05] expose l'évolution des OLPT vers les entrepôts de prochaine génération. Le lecteur pourra consulter la synthèse de ces écrits à l'annexe E.

Plus on intègre de processus et de données à l'entrepôt, plus son volume devient astronomique. Le volume constitue donc un problème sur lequel Inmon se penche. Il nomme DW2.0 les entrepôts de données de prochaine génération. Pour mieux gérer le volume de données, il propose un modèle d'archivage des données en quatre parties.

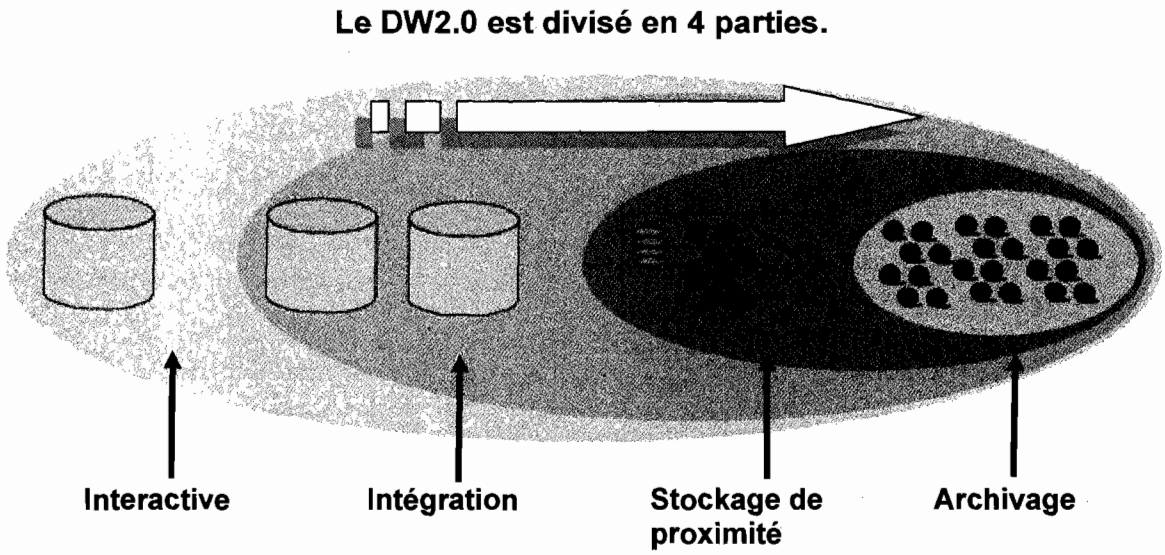


Figure 3.24 Le DW2.0 d'Inmon.

La figure 3.24 résume la vision d'Inmon. Les données sont intégrées à l'entrepôt de la partie «interactive» vers la partie «intégration». Par la suite, elles progressent de l'état «courant» dans la partie intégration vers l'état «latente» de la partie «stockage de proximité» à partir de laquelle les données sont encore disponibles pour l'extraction. Après une fréquence prédéterminée, les données migrent vers l'archivage physique sur support. Les données ne sont plus accessibles directement pour l'extraction. Les données archivées pourront être réintégrées à l'entrepôt au besoin en repassant en sens inverse, de la partie archivage à la partie stockage de proximité et au besoin à la partie intégration.

Le lecteur pourra consulter les détails de la présentation de Inmon qui décrit toutes les particularités de chaque phase sur le site Internet : http://www.inmoncif.com .

Partie 2
Méthodologie

Cette deuxième partie permet de justifier les choix de conception qui seront utilisés dans la partie 3. Elle permettra aussi d'évaluer certains outils existants et d'en faire une recommandation d'achat. À la fin de ce chapitre, une méthodologie sera adoptée pour la suite du document afin de répondre aux besoins de l'UQTR. Elle sera présentée en deux chapitres, soient :

4. Analyse

5. ETC

Chapitre 4
Analyse

L'objectif du mémoire découle de la mise en place initiale de chaque étape de création d'un entrepôt de données. Ce chapitre précise les choix relativement à la méthodologie qui sera mise de l'avant au chapitre 6. Ce chapitre est divisé comme suit :

4.1 Justification des choix

Découlant du chapitre 3, il convient maintenant de faire les meilleurs choix et de façon éclairée pour chaque phase du développement.

4.1.1 Approche de base

Après examen, l'approche par *data marts* indépendants est rejetée. Cette approche fortement orientée-sujet personnalise les analyses rendant ainsi difficile une analyse croisée fiable. De plus, la répétition des traitements et des données rend très lourde la tâche d'entretien des processus.

Il faut maintenant comparer les deux autres options. Deux grands concepts, présentés par deux auteurs, s'opposent : Inmon et Ralph Kimball. La différence entre les deux méthodologies peut être décrite comme la distinction entre l'extensibilité et la puissance versus la rapidité et la simplicité. L'approche d'Inmon [Inmon 96] exige la création d'un modèle d'entrepôt de données complet dans un premier temps. Le résultat de ce processus peut alors être employé dans les étapes suivantes comme base pour modeler les extraits dimensionnels et non dimensionnels. Dans l'approche de Kimball [Kimball et al. 05], des structures dimensionnelles sont créées sans avoir besoin de créer entièrement l'entrepôt. Si les structures dimensionnelles correspondent aux exigences de ce qu'une organisation exige pour accomplir ses besoins d'analyse de données, alors l'approche de Kimball est une manière plus rapide et plus simple pour créer un entrepôt de données.

Cependant, si d'autres types de magasins analytiques de données sont nécessaires en plus des structures dimensionnelles, l'approche d'Inmon propose une méthode plus puissante. D'autre part, le modèle dimensionnel est développé strictement pour l'analyse d'OLAP en mode utilisateur.

Pour faire notre choix final, il faut considérer que l'on veut unifier les données. Il faut partager les données de l'admission avec celles du Service des finances et des autres systèmes transactionnels de l'UQTR. Il est donc important que le modèle de données soit intuitif pour les utilisateurs finaux. Un premier *data mart* doit être fonctionnel rapidement et tous les processus d'affaires devront être intégrés à tour de rôle. Il est essentiel d'obtenir les résultats, et ce, dans les meilleurs délais. Pour ces raisons, le choix qui s'impose est le modèle dimensionnel «*bottom-up*» de Kimball.

4.1.2 Approche de base «orientée»

L'approche de base privilégiée «bottom-up» oriente la méthodologie du projet dans son ensemble. Il peut être très coûteux de la modifier par la suite. L'approche de base «orientée», correspondant à l'approche de base de «Kimball» à l'étape 4.1.1, est l'approche «piloter par les besoins utilisateurs».

L'approche «piloter par les besoins utilisateurs» est risquée selon [LIST et al. 02]. Ils conseillent de choisir une approche complémentaire. Si un gestionnaire attitré à un poste est remplacé par un autre gestionnaire, les besoins de ce dernier changeront nécessairement. Les besoins modélisés par le premier gestionnaire pourraient devoir être modifiés complètement. C'est pourquoi, nous allons fusionner les approches «besoins» et «objectifs».

Un questionnaire à l'intention des gestionnaires a été conçu pour pallier à la faiblesse de l'approche par besoins utilisateurs. Ce questionnaire permet de décrire les objectifs globaux du processus d'affaires étudié en fonction des objectifs globaux de l'entreprise. Le gestionnaire, avant d'énoncer ses besoins doit alors se repositionner face aux objectifs de son service et ceux globaux de l'entreprise. L'approche «piloter par les objectifs» sera donc la première à exécuter avant de poursuivre avec l'approche «piloter par les besoins utilisateurs». Cette dernière approche comporte plusieurs dimensions et plusieurs tables de faits. Les utilisateurs seront sollicités afin de participer à la définition de leur processus d'affaires.

4.1.3 Type de serveur OLAP

Pour nos besoins en fonction de l'infrastructure de l'UQTR, le modèle d'architecture ROLAP sera retenu. Ce modèle est plus flexible et extensible en fonction des changements de l'entreprise. On pourra plus facilement et plus rapidement ajuster un modèle de données de l'entrepôt.

Un autre avantage du type de serveur ROLAP nous permettra de répondre plus facilement au besoin de *data mining* dans le futur. À court terme, il sera possible d'explorer les données de l'entrepôt avec les outils existants pour le modèle relationnel, fait à ne pas négliger. Il est plus facile de convertir une architecture ROLAP vers le MOLAP que l'inverse.

Un dernier argument militant en faveur de l'expertise actuelle du Service est plutôt orienté relationnel. S'il faut rapidement mettre en place l'entrepôt pour un premier tableau de bord,

ce choix nous facilitera la tâche. Nous pouvons conclure que la souplesse du modèle ROLAP nous est favorable.

4.2 Le cycle de vie décisionnel

Si l'on compare la méthodologie X-Meta à celle de Kimball, cette dernière est plus détaillée et plus de phases différentes sont couvertes par cette méthode. La méthode de développement du prototype de X-Meta correspond à l'axe des données de Kimball.

De plus, le cycle de vie décisionnel de Kimball est expliqué en détail, étape par étape, dans son volume «*Guide de conduite de projet : Entrepôt de données.*»

Le choix retenu est donc, pour ces motifs, le cycle de vie décisionnel de Kimball.

4.3 L'architecture logique

Considérant les choix précédents et que l'on veut unifier les données de l'entrepôt, il faut utiliser les dimensions et les faits conformes, voici le type d'architecture logique de la solution : dimensionnel «bottom-up» avec *staging area* (zone tampon de prétraitement des données). La zone de données temporaire est représentée à la figure 4.1 par le CDC. Même si le schéma des données est dimensionnel, la structure des données est emmagasinée dans une base de données relationnelle Oracle qui permettra d'intégrer d'autres outils d'extraction via une interface (API) afin de présenter les données sous d'autres formes que le cube de données.

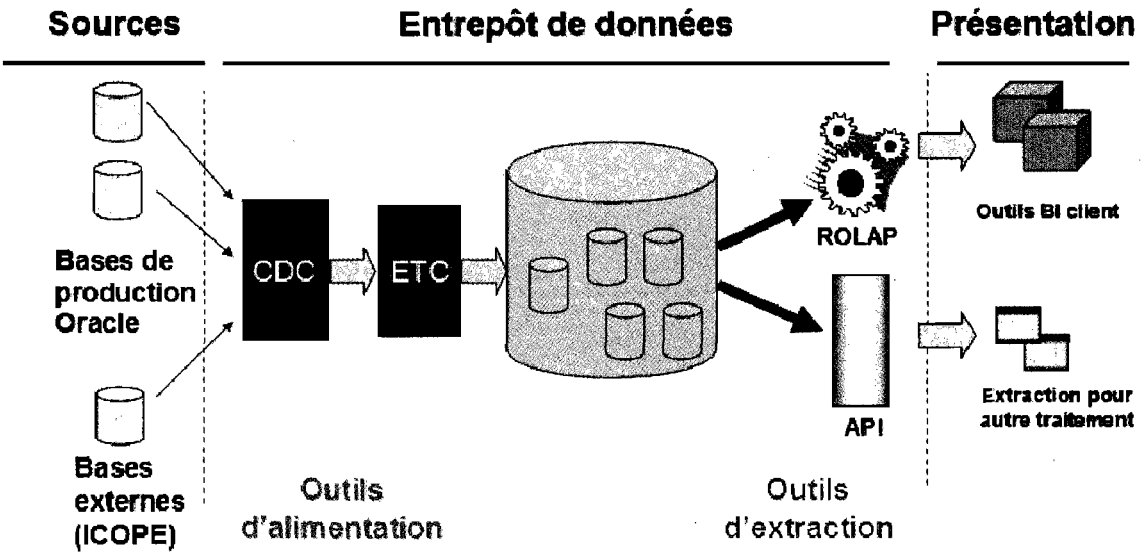


Figure 4.1 Architecture logique de l'entrepôt de données de l'UQTR.

4.4 L'architecture physique

Bien que la taille de l'UQTR soit considérée comme relativement petite, il est recommandé à des fins de performance que l'instance de l'entrepôt de données ainsi que ses services soient sur un seul serveur comme le montre la figure 4.2. Les concepteurs d'outils le recommandent pour la simple et bonne raison que certaines analyses gourmandes diminueraient la performance des systèmes transactionnels, ce qu'il faut éviter.

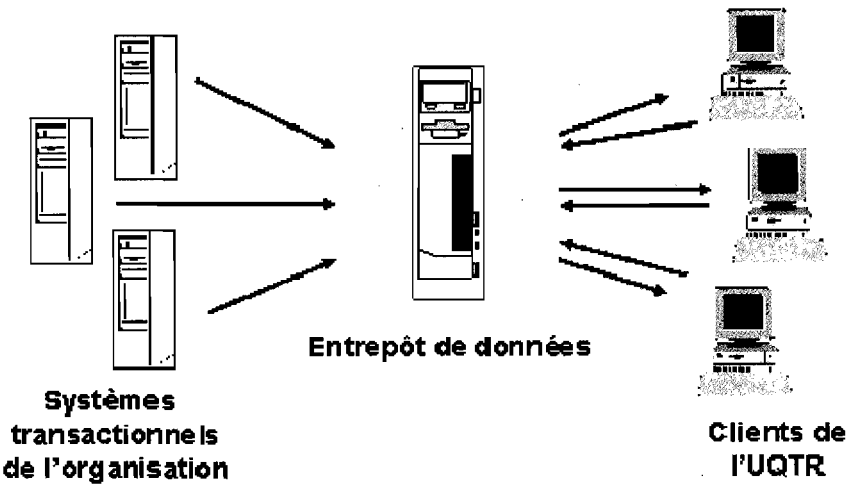


Figure 4.2 Architecture physique de l'entrepôt de données de l'UQTR.

4.5 Évaluation des besoins des utilisateurs

Afin de pallier à la faiblesse de l'approche de base «orientée besoins utilisateurs», un questionnaire a été réalisé afin de recueillir les objectifs d'affaires des futurs utilisateurs de l'entrepôt de données. Ce questionnaire tente d'identifier : «Quelle est la perspective du décideur et des connaissances diverses dont il dispose ?» Le lecteur pourra consulter le formulaire à l'annexe F. Monsieur Rémy Auclair a rempli le questionnaire pour l'étude en cours.

L'objectif de ce questionnaire est de comprendre l'activité et les objectifs des utilisateurs afin de les traduire en besoins, c'est-à-dire en de termes de données et d'analyse. L'avis de l'utilisateur est indispensable pour définir les exigences du projet. Ce questionnaire est composé de quatre parties :

Partie 1 : Identification
Partie 2 : Utilisation des systèmes de gestion de l'UQTR
Partie 3 : Utilisation d'autres sources de données
Partie 4 : Prise de décision et analyse

4.6 Les métadonnées

«Il est nécessaire de privilégier un langage commun entre les différents services d'une entreprise pour partager une vision identique de la notion de donnée où l'aspect de partage de savoir est fondamental. Cela passe notamment par l'organisation d'ateliers de partage de savoir et/ou par l'élaboration incrémentale d'un dictionnaire de données d'entreprise».

Nicolas Debaes, Architecte Senior (Octo Technology)

Les métadonnées représentent le dictionnaire unique de l'entrepôt. Lorsqu'une donnée est intégrée à l'entrepôt, si l'on veut suivre sa trace, il faut être en mesure de connaître sa provenance, sa définition technique et celle qui sera présentée à utilisateur. D'autres mesures ou informations peuvent être conservées. Il convient de proposer une liste de certaines métadonnées.

Une carte de localisation 2D permettrait d'évaluer l'impact d'une modification d'une donnée sur l'ensemble des processus. Si une donnée des OLTP est modifiée, il faut savoir où intervenir dans le processus d'intégration de l'entrepôt. Certains logiciels ETC permettent d'obtenir cette carte de localisation. Si l'on veut mettre en place cette fonctionnalité, il faut savoir qu'une donnée intégrée provient d'une seule source. Cependant, précisons qu'une source peut alimenter plus d'une donnée cible dans l'entrepôt.

La liste des transformations permet d'associer à une table de dimensions ou de faits le ou les processus qui l'alimentent. Il suffit de conserver les processus associés à une table de l'entrepôt dans une autre table (Ex. : EID_MD_TRANSFORMATION). Chaque processus décrit pourrait avoir les caractéristiques suivantes : un état (être actif ou inactif), une fréquence de chargement, un indicateur d'utilisation CPU.

Chaque processus de transformation ETC est nommé et ce nom doit être unique. L'administrateur de l'entrepôt doit être informé si un processus est arrêté avant sa fin normale. Il faut être en mesure de consulter la liste des chargements des processus de l'entrepôt avec leur état (SUCCÈS, ÉCHEC). Si un processus s'exécute et son état devient «ÉCHEC», un courriel sera automatiquement expédié à l'administrateur. Une table alimentée par chaque processus déterminera les processus démarrés, l'heure, et sa finalité (Ex. : EID_CHARGEMENT).

La plupart des outils ETC offrent au minimum un dictionnaire de données des tables et des champs des processus, mais la fonctionnalité se limite aux noms techniques. Il est important dans un processus d'unification des données que les différents concepts du domaine soient nommés et expliqués à l'utilisateur. Pour ce faire, une définition d'affaires des concepts de l'entreprise sera conservée dans deux tables (Ex. : EID_MD_TABLES, EID_MD_CHAMPS). Finalement, pour chaque table et champ, il faut définir les règles d'accès, soit par rôle ou par usager. Le tableau 4.1 nous résume la liste des métadonnées proposées.

Tableau 4.1
Liste des métadonnées proposées

Description des métadonnées	Nom de table associée	No	Particularités
Liste des transformations	EID_MD_TRANSFORMATION	1	IND_CPU (indicateur d'utilisation CPU)
		2	IND_ACTIF (processus actif ou non)
Liste des chargements	EID_CHARGEMENT	3	FREQ_CHARGEMENT (fréquence de chargement du processus)
		4	Indique si le processus de chargement s'est exécuté avec succès ou non
Dictionnaire des données (table et champ)	EID_MD_TABLES	5	NOM_CHAMP (définition technique)
		6	DESC_CONCEPT_DATA (définition d'affaires des données)
		7	DT_BEDUT (historisation des structures)
		8	DT_FIN (historisation des structures)
	EID_MD_CHAMPS	9	NOM_CHAMP (définition technique)
		10	DESC_CONCEPT_TB (définition d'affaires des concepts)
		11	DT_BEDUT (historisation des structures)
		12	DT_FIN (historisation des structures)
Règles d'accès et de sécurité	EID_MD_ACCESS_DATA	13	rôle ou usager
	EID_MD_ACCESS_TB	14	rôle ou usager

Même si un logiciel ETC permet de suivre certaines métadonnées, il peut être plus sécuritaire d'en faire la gestion dans une étape antérieure. C'est pourquoi le développement d'un outil pour la gestion des métadonnées est proposé au chapitre 6.

4.6.1 Modification des structures des systèmes transactionnels

L'entrepôt de données se doit de rendre une information juste et exacte puisque le résultat présenté aux demandeurs doit être toujours le même peu importe le moment de la demande.

Lors de la mise en place du prototype, des modifications aux structures des tables des systèmes transactionnels ont provoqué des erreurs d'intégration des données.

L'administrateur de l'entrepôt se doit d'être informé des changements de structures et il décidera par la suite si une modification du côté de l'entrepôt s'impose.

Pour pallier à ce problème, nous avons développé le logiciel OAD (Outil d'analyse des DDL). Ce système, à partir des journaux de recouvrement «*redo logs*» d'Oracle, capte toutes les commandes de modification de structures (DDL) des systèmes transactionnels désirés. Une commande DDL peut être par exemple la modification d'un champ à la structure d'une table de la base de donnée avec la commande «ALTER TABLE». L'administrateur de l'entrepôt peut suivre à la trace ces modifications et agir de façon ponctuelle et journalière. Pour permettre l'accès aux «logs» d'Oracle, le logiciel LOGMINER d'Oracle fut installé sur la base de données. Ce logiciel permet d'avoir accès aux journaux (*logs*) des transactions archivées et des transactions toujours en mémoire vive du serveur de base de données. Ces journaux contiennent toutes les transactions de la base de données permettant le recouvrement entier de la base de données en cas de bris matériel ou logiciel ou encore d'un sinistre.

4.6.2 Historisation de la structure des données

En examinant les données de certains systèmes de l'UQTR, on s'aperçoit qu'il y a des valeurs manquantes dans certaines tables. Comment expliquer ce phénomène ? La raison principale est qu'afin d'améliorer les processus transactionnels, des ajouts ou des modifications ont été apportés aux structures existantes à un moment dans le temps. Certains systèmes de l'UQTR ont des données depuis la création soit l'année 1969. Il est certain que le développement des systèmes d'information a modifié les structures originales. Il y a eu de nouveaux champs qui ont été créés.

Si la date de diplômation a été ajoutée en 1980, aucune donnée n'est présente dans tous les enregistrements précédents 1980. Si un demandeur veut calculer la durée moyenne des études de tous les étudiants depuis le début, puisque la valeur «NULL» est contenue dans certains enregistrements, le résultat de cette moyenne sera donc biaisé. Il faut donc porter une attention particulière aux processus de remplacement des valeurs manquantes lors des prétraitements.

Afin de suivre l'évolution des changements des structures de données dans le temps et d'en informer le demandeur afin qu'il comprenne ses résultats, il faut suivre les modifications de structures dans le temps. Pour améliorer l'historisation des structures de [EDER-KONCILIA 02], nous proposons une historisation complète des structures identiques à l'historisation des données. Un champ «DT_DEBUT» et un champ «DT_FIN» permettront de suivre les modifications des structures dans le temps. Cette solution permet de résoudre le problème de changement des structures dans le temps.

Les outils de chargement ou de présentation des données existants ne répondent pas entièrement à la méthodologie proposée, omettant complètement l'historisation des structures de données et laissant ainsi les résultats obtenus avec des biais plus ou moins cruciaux à la prise de décision. Il faudrait trouver une façon d'intégrer l'information aux outils existants sans être obligé de développer entièrement nos outils.

4.7 Modélisation architecturale des données

La modélisation des données est une étape cruciale au développement. C'est dans cette étape que l'on schématise le processus d'affaires d'un sujet donné. Il faut donc avoir un processus d'affaires déterminé pour le sujet d'affaires que l'on désire intégrer à l'entrepôt de données.

Il y a trois méthodes proposées. La première, la méthode à trois niveaux qui est celle utilisée lors de la modélisation des systèmes transactionnels. Elle est composée du modèle conceptuel, du modèle logique et du modèle physique. Puisque cette méthode omet complètement la modélisation de l'aspect dimensionnel requis dans notre cas et puisque le choix de l'architecture logique est «dimensionnelle «bottom-up» avec dimensions et faits conformes de Kimball», cette méthode ne sera pas retenue.

Les autres méthodes proposent une modélisation objet de l'aspect dimensionnel. Il s'agit de la méthode à quatre niveaux avec l'arbre du sujet [SHUNUNG et al. 05] et la méthode à quatre niveaux avec DMF [GOLFARELLI-RIZZI 98]. Ces méthodes sont composées du modèle conceptuel, du modèle logique, du modèle objet et du modèle physique.

Quoique la méthode avec DMF soit intéressante, elle est moins intuitive que la méthode avec l'arbre du sujet. Cette dernière est retenue puisque on peut facilement visualiser tout le contexte dimensionnel. Nous ajoutons cependant une légère modification visuelle à cette méthode à la figure 4.3 : chaque type d'objet (sujet, fait, mesure, ...) aurait son propre symbole.

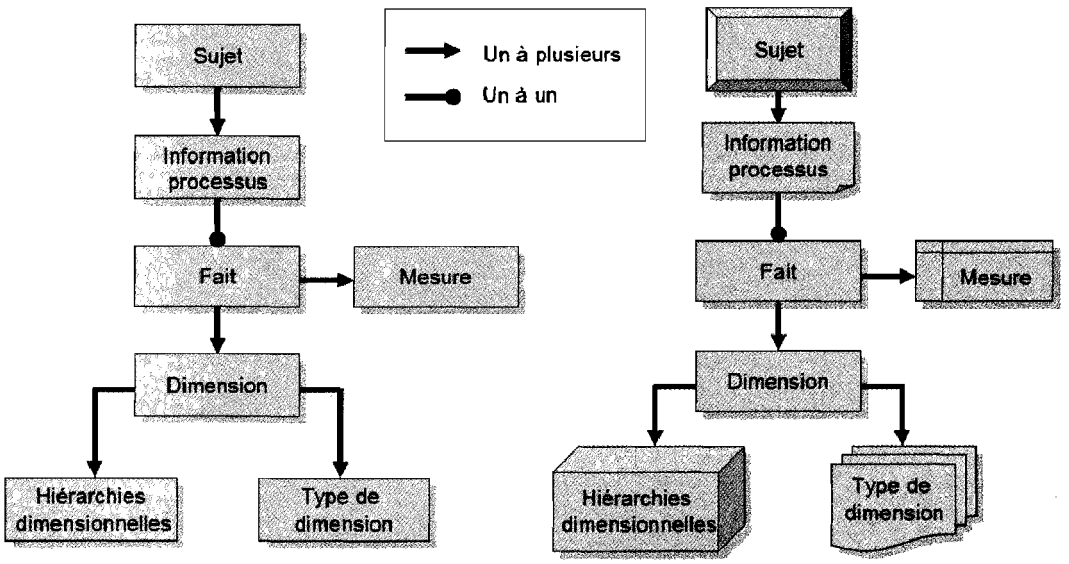


Figure 4.3 Arbre de sujets avec nouveau symbolisme (à droite).

4.8 Proposition d'une méthode de conception

"En théorie, la pratique et la théorie sont identiques. En pratique, elles sont différentes."

Cette pensée écrite au bas du courriel d'un de mes collègues m'a toujours fait sourire. En théorie, les données des systèmes sont intégrées par un outil ETC vers une zone temporaire optionnelle et ensuite vers l'entrepôt de données pour finalement être disponibles par des services de présentation. En pratique, il manque techniquement beaucoup d'étapes avant que l'outil ETC puisse débuter ses opérations.

Une étape importante doit donc être incluse lors de la modélisation objet, troisième étape de la modélisation architecturale des données. Il faut identifier si le modèle étoile sera de type temporel ou non. S'il n'est pas temporel, les vues matérialisées pourront être utilisées pour alimenter l'ETC. Sinon, le processus est tout autre. Il faut gérer le temporel et détecter seulement dans les enregistrements modifiés les données pour lesquelles une trace

temporelle est requise. Par exemple, je ne veux pas suivre la trace de toutes les modifications du champ téléphone d'un étudiant. Par contre dans le même enregistrement, si la date de naissance est modifiée, il faut la détecter et faire passer cette modification à l'entrepôt. La composante CDC (change data capture) est alors requise à cette fin.

Il restera, dans la partie conceptuelle, à définir toutes les étapes nécessaires à la préparation des données avant d'intégrer les données par les outils ETC à l'entrepôt.

4.8.1 Vues matérialisées

L'exploration des vues matérialisées (vues converties en table physique) a été très intéressante. Son implantation était simple et rapide tout en permettant de voir séquentiellement toutes les modifications sur les données des tables sources pointées. La conception et l'implantation d'un prototype ont été faites avec les vues matérialisées.

Le principe est, qu'à chaque vue, est associée une table de *logs* conservant les modifications de la table source sur laquelle la vue est liée. Dans la table de *log*s de la vue (liste des modifications de tous les enregistrements), un indicateur nous informe de l'opération faite sur l'enregistrement. Cet indicateur peut être 'I' (insert), 'U' (update) ou 'D' (delete). Il faut donc traiter différemment l'enregistrement en fonction de son état. Toutes les modifications s'accumulent dans la table de *logs* jusqu'à ce qu'une demande (manuelle ou automatique) intervienne et rafraîchisse la vue avec les nouvelles modifications. Presque au même moment, une fois la vue rafraîchie, la table de *logs* se vide et la vue matérialisée en table physique est alors à ce moment précis seulement identique à la table source du système transactionnel.

Si l'utilisateur n'a pas besoin de l'aspect temporel des données, les vues s'avèrent très utiles. Par contre si la trace des modifications dans le temps est requise, les vues matérialisées ne permettent pas de suivre l'historisation en temps réel des données. Les données ne sont pas accessibles directement dans le *logs* de la vue mais encodées dans un vecteur. On ne peut donc pas savoir quel champ est modifié. Une perte d'information peut survenir entre le traitement de mise à jour de l'entrepôt et une nouvelle modification dans la seconde. Aux fins de notre projet, on ne peut se permettre de perdre une donnée.

4.8.2 CDC

Voici une brève explication du CDC d'Oracle (Change Data Capture). À l'aide de déclencheurs (triggers), Oracle capture les ordres DML et les insère dans une table de modifications (change table) afin que les changements puissent être publiés ultérieurement (et seulement les changements). (Source : developpez.com)

L'outil CDC permet aussi de gérer les données non temporelles, c'est pourquoi notre choix s'est arrêté sur cette façon de faire. De plus, toute l'information des données modifiées (Ex. : **:old.nom** et **:new.nom**) est visible textuellement et donc traitable par un autre processus.

Cet outil est présentement fonctionnel sur la base de données transactionnelle de l'UQTR et capte les modifications des tables de l'admission. Il nous reste seulement à le brancher à l'outil ETC qui sera choisi lors de la recommandation d'achat au chapitre 5.

4.8.3 Méthodologie proposée

À chacun des points précédents du chapitre 4, une justification des choix permet de résumer la méthodologie adoptée. Il convient à ce moment-ci de faire une synthèse des étapes qui en découlent ainsi que des choix retenus. Le tableau 4.2 présente cette synthèse.

Tableau 4.2

Synthèse de la méthodologie proposée et des choix possibles

	Étapes	Choix possibles pour chaque étape	Choix retenus
1	**Approche de base**	☐ «*top-down*» (Inmon) ☐ «*bottom-up*» (Kimball) ☐ Hybride	**«*bottom-up*»** (Kimball)
2	**Approche «orientée»**	☐ Piloter par les données ☐ Piloter par les besoins utilisateurs ☐ Piloter par les objectifs ☐ Hybride	**Hybride** (objectifs et besoins)
3	**Architecture logique**	☐ Entité-relation «*top-down*» ☐ Dimensionnel «*bottom-up*» (sans *staging area permanent*) ☐ Dimensionnel «*bottom-up*» avec dimension et faits conformes (avec *staging area non perm.*) ☐ *Data mart* independent	**Dimensionnel «*bottom-up*» avec dimension et faits conformes** (avec *staging area non permanent*)
4	**Cycle de vie décisionnel**	☐ X-Meta ☐ Méthode Kimball	Méthode Kimball
Une fois les 4 étapes choisies, effectuer les choix des étapes 5 et 6.			
5	**Architecture physique**	☐ ROLAP ☐ MOLAP ☐ HOLAP	**ROLAP**
6	**Modèle architectural des données**	☐ À 3 niveaux ☐ À 4 niveaux avec l'arbre du sujet[1] ☐ À 4 niveaux avec l'arbre du sujet[2] ☐ À 4 niveaux avec DMF	**À 4 niveaux** (avec l'arbre du sujet[2])

La méthodologie représente l'ensemble des choix retenus pour la mise en place d'un entrepôt de données.

Après ces choix, la conception s'amorcera et sera basée sur cette méthodologie en appliquant concrètement le cycle de vie décisionnel et ses processus itératifs.

[1] Modèle à 4 niveaux avec l'arbre du sujet de [SHUNUNG et al. 05]
[2] Modèle à 4 niveaux avec l'arbre du sujet de [SHUNUNG et al. 05] avec nouveau symbolisme

Chapitre 5
ETC (Extraction, Transformation et Chargement)

Ce chapitre fait le tour des principaux outils ETC. Il introduit les notions d'extraction, de transformation et de chargement. De plus, sera introduit la phase essentielle à toute intégration : celle de «la préparation des données». Le point «Évaluation des outils existants» a été réalisé en collaboration avec mon collègue au projet, monsieur Michel Charest. Il servira de base pour une formulation d'une première ébauche de recommandation d'achat pour la mise en place de l'entrepôt institutionnel de l'UQTR. Ce chapitre est divisé en trois points :

5.1 ETC

Qu'est-ce qu'un ETC (ETL en anglais). Selon Wikipédia :

« Extract-Transform-Load » est connu sous le terme ETC (ou parfois : datapumping). Il s'agit d'une technologie informatique intergicielle (middleware) permettant d'effectuer des synchronisations massives d'information d'une banque de données vers une autre. Selon le contexte, on traduira par « alimentation », « extraction », « transformation », « constitution » ou « conversion », souvent combinés. Source : Wikipédia

On peut retrouver dans la littérature le terme ETC ou ECT. Et non, ce n'est pas seulement une inversion de lettre. Dans les deux cas, il s'agit bien d'extraire, de transformer et de charger l'entrepôt de données. L'inversion indique une méthode d'alimentation différente.

La méthode ETC est l'approche traditionnelle pour intégrer les données à l'entrepôt de données. C'est la méthode la plus répandue actuellement. Elle consiste à installer un «moteur ETC» du côté serveur par lequel toutes les transformations seront effectuées. La plupart des solutions commerciales fonctionnent de cette façon notamment : COGNOS, Informatica, IMB.

L'approche ECT (ELT en anglais) consiste plutôt à décentraliser le traitement de chargement vers les sources hétérogènes. Chaque base de données impliquée va s'occuper de charger ses données vers le conteneur central, l'entrepôt, mais plus particulièrement dans une zone intermédiaire de prétraitement le «staging area». C'est seulement après le chargement des données à l'entrepôt que les transformations vont s'amorcer. Les compagnies Genio et Sunopsis offrent ce type d'ECT.

Il faut mentionner que Business Object avec son outil «Data Integrator» offre les deux approches. Dans cette section, l'approche ETC sera considérée.

Souvent associée à l'outil ETC, on entendra parler de zone temporaire de prétraitement traduit par «staging area». Certains développements n'auront pas besoin de cette zone puisque les données transitent sans modification majeure. Ils n'utilisent pas comme données d'entrée, les données de sortie d'un autre processus. Ils n'attendent donc pas la réussite du processus précédent avant de poursuivre. Des projets de très petites tailles peuvent être réalisés sans «staging area». Pour d'autres projets, si le but était seulement de libérer des traitements analytiques les systèmes transactionnels, les structures des OLTP seraient presque les mêmes que celles de l'entrepôt.

Dans d'autres cas, et c'est celui de l'UQTR, beaucoup trop de données sont à traiter. Pensons aux journées d'inscriptions où les systèmes transactionnels travaillent sans relâche. Il serait navrant de perdre des données lors de ces périodes de pointe. Il existe deux types de «staging area» : un temporaire et l'autre permanent ("persistant staging area"). Le choix de l'outil ETC est en lien étroit avec le type de «staging area» que l'on pourra utiliser.

On utilise le «staging area» temporaire dans le but de diminuer les impacts des chargements sur les OLTP. On change les données dans une zone, on libère les OLTP et après on effectue les transformations pour ensuite les intégrer à l'entrepôt. Lorsque les données sont transférées dans la zone temporaire, nous effectuons le 'E' de ETC, c.-à-d. la phase d'extraction. Lorsque les transformations s'exécutent, c'est le 'T' de l'ETC qui se met en branle. Finalement, le chargement des données transformées et nettoyées se fait chargées vers l'entrepôt. C'est la lettre 'C' de ETC qui est en action. Normalement, la fenêtre d'extraction doit être telle que si un chargement échoue, il ne faut pas extraire de nouveau les informations des OLTP pour recommencer le travail. C'est donc la fréquence de chargement qui peut déterminer la grandeur de la fenêtre d'extraction. Cette zone temporaire peut être une copie de la base de données de production et sert à emmagasiner les données en attendant de les transformer.

Le «staging area» permanent est l'endroit où sont emmagasinées les tables après transformation. Kimball le nomme «persistent staging area» et Inmon le nomme «entreprise datawarehouse». C'est dans ce «staging area» que l'on retrouve les schémas étoiles ou flocons de neige et que le remplacement des clés primaires des OLTP se substitue aux nouvelles clés de l'entrepôt. Ce «staging area» gère l'évolution lente (de type 1,2 ou 3) ou rapide des dimensions. Les traitements résultants sont soient agrégés ou unitaires.

L'évolution lente permet de suivre la trace des modifications dans le temps. Elle sert à détecter toutes les modifications et à conserver l'historique des données. La plupart des outils ETC offre les trois types d'évolution des dimensions. Le «type 1» écrase tout simplement l'ancienne donnée de l'entrepôt par sa nouvelle. Le «type 2» insère en tout temps. Pour chaque modification il y aura une nouvelle ligne par enregistrement modifié. Le «type 3» ajoute une nouvelle colonne conservant uniquement les dernières valeurs «old» et «new» d'un même champ. Même si je modifie quinze fois les données, il n'y aura que les deux dernières valeurs de conservées soit : la nouvelle valeur dans la colonne «new» et la dernière valeur dans la colonne «old».

5.1.1 Préparation des données

Plan de projet ETC (proposé par Rémy Choquet – Université Lyon 2) : «L'ETC dans le cadre de l'entreposage de données».

Monsieur Choquet propose une méthodologie de construction des tâches ETC qui se divise en huit phases. On entend par tâche, un processus ETC complet. Cette méthode pose le cadre de la tâche d'une manière générale. Il conseille d'utiliser une approche pas à pas si une tâche est composée de plusieurs autres. Il s'inspire des 38 sous-systèmes ETC de Kimball et de ses 10 étapes en les synthétisant en 8 groupes. Le tableau 5.1 identifie le responsable de la tâche à accomplir.

Tableau 5.1

Les différents responsables des processus ETC

DBA	***Data base administrator* (administrateur de la bd)**
A-Système	**Analyste système (systèmes transactionnels)**
A-ETC	**Administrateur de l'outil ETC**
G-ETC	**Gestionnaire de l'outil ETC**
D-ETC	**Développeur de l'outil ETC**
S-Q-D	**Service de la qualité de données**

L'ensemble de ces phases correspond à la phase de préparation des données (figure 5.1). Elle se situe en aval des systèmes OLTP jusqu'en amont du chargement des données vers l'entrepôt.

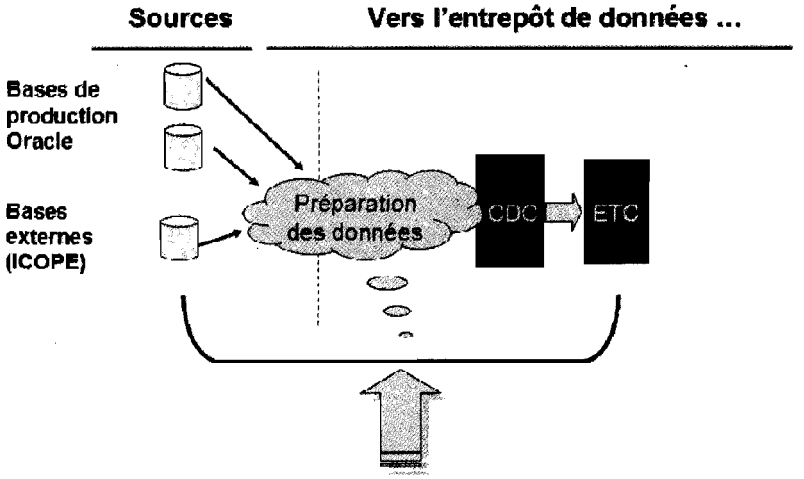


Figure 5.1 Zone de préparation des données.

5.1.2 Phases d'intégration avec l'ETC

Les trois premières phases, présentées dans le tableau 5.2, complètent la phase de préparation des données. On met en place les outils de développement et on énonce le processus d'affaires étudié avec les utilisateurs concernés. Dans un premier temps, il faut s'assurer d'avoir les données que l'on veut analyser. Si les données ne sont pas présentes dans les systèmes sources, on ne peut ni les inventer ni les extraire. Par la suite, il faut s'assurer de la qualité de ces données. Ce processus est assez complexe. Il faut vérifier les incohérences dans la base de données et vérifier si les données ont bien le sens que les utilisateurs leur donnent. L'analyse de la qualité des données des systèmes sources peut être un point déterminant dans la poursuite ou non d'un projet.

La première phase est assez évidente puisque sa description parle d'elle-même. La phase deux explore les besoins du processus d'affaires étudié, trouve l'emplacement des données dans les systèmes sources et analyse ses données. La phase trois permet concrètement d'amorcer l'analyse des systèmes sources et de créer la table de correspondance des données sources vers les données cibles.

Tableau 5.2

Les phases du processus ETC (1 à 3)

Phase	Description de la phase	No	Description de la tâche	Responsable
I	**Mise en place de l'environnement de développement**	**1**	Configurer l'infrastructure matérielle.	**DBA**
		2	Installation des logiciels et outils.	**DBA / A-ETC**
		3	Mettre en place les documents sur les meilleures pratiques.	**G-ETC / A-ETC**
II	**Analyse des besoins métiers**	**1**	Revue de la documentation existante entre source et cible.	**A-ETC / A-Système**
		2	Définition et documentation des règles métier.	**A-ETC / A-Système**
		3	Analyse des systèmes sources.	**A-ETC / A-Système**
		4	Définition de la portée des phases de projet.	**G-ETC**
III	**La conception des mises en correspondance des données**	**1**	Revue du modèle de données de l'entrepôt de données.	**A-ETC**
		2	Revue des règles métier.	**A-ETC**
		3	Analyse des systèmes sources.	**A-ETC**
		4	Création du document de mise en correspondance des données.	**A-ETC**

La phase quatre permet de valider les données. Lorsque des anomalies sont trouvées, le gestionnaire les mentionne aux utilisateurs. La phase cinq permet la mise en place du modèle. Le tableau 5.3 décrit les phases IV et V du processus ETC.

Tableau 5.3

Les phases du processus ETC (4 et 5)

Phase	Description de la phase	No	Description de la tâche	Responsable
IV	**Stratégie de qualité des données**	1	Définition des règles de qualité des données.	**G-ETC / S-Q-D**
		2	Documentation des défauts de donnés.	**G-ETC / S-Q-D**
		3	Affectation de la responsabilité des défauts de données.	**G-ETC / S-Q-D**
		4	Création du document de mise en correspondance des données.	**G-ETC / S-Q-D**
		5	Sensibilisation des utilisateurs finaux des défauts des données.	**G-ETC / S-Q-D**
		6	Intégration des règles de qualité dans le document de mise en correspondance.	**G-ETC / S-Q-D**
V	**Développement des processus ETC**	1	Revue du document de mise en correspondance.	**D-ETC**
		2	Développement des dimensions simples.	**D-ETC**
		3	Développement des dimensions SCD-2 (historique).	**D-ETC**
		4	Développement des dimensions SCD-2 (incrémental).	**D-ETC**
		5	Développement des tables de faits (historique).	**D-ETC**
		6	Développement des tables de faits (incrémental).	**D-ETC**
		7	Automatisation des processus.	**D-ETC**

La phase six permet le contrôle de qualité. Tous les tests sont faits une fois le premier processus ETC implanté.

Une fois la phase six réussie, le déploiement s'amorcera à la phase sept et les données seront intégrées à l'entrepôt. La dernière phase permet une rétroaction des problèmes rencontrés afin d'en limiter la reproduction dans l'avenir et de laisser une trace pour quelqu'un qui cherche une solution. Le tableau six décrit les phases VI, VII et VIII du processus ETC.

Tableau 5.4

Les phases du processus ETC (6 à 8)

Phase	Description de la phase	No	Description de la tâche	Responsable
VI	Tests unitaires / Tests d'assurance qualité / Tests d'acceptation	1	Mise en place de l'environnement de test.	DBA / A-ETC
		2	Création des plans de tests et les scripts.	A-Système
		3	Chargement des données.	D-ETC
		4	Exécution des scripts de tests unitaires.	A-Système
		5	Contrôle de la qualité des données.	A-Système
		6	Validation des données.	A-Système
		7	Validation des règles métier.	A-Système
		8	Obtention de l'acceptation.	G-ETC
VII	Déploiement	1	Création des documents de support.	A-ETC
		2	Création des documents des mécanismes de récupération.	A-ETC
		3	Mise en place de l'environnement de production.	A-ETC
		4	Chargement des données historiques.	A-ETC
		5	Ordonnancement des processus incrémentaux.	A-ETC
VIII	Maintenance	1	Développement des rapports d'audit pour les problèmes connus.	A-ETC
		2	Vérification des journaux d'exécution.	A-ETC
		3	Mise en place de l'environnement de production.	A-ETC

5.2 Évaluation des outils existants

Plusieurs compagnies offrent des suites intégrées pour la gestion des entrepôts de données. «Suites intégrées» veut dire qu'ils offrent à la fois l'outil ETC et l'outil de présentation des données. C'est la principale raison pour laquelle il n'est pas évident de séparer les deux outils dans l'analyse. Certains fabricants offrent l'un ou l'autre de ces outils. Dans la littérature, ce genre d'évaluation est inexistant. Nous allons donc innover et faire profiter nos pairs, les autres universités du réseau, de cette évaluation.

5.2.1 Critères pour les outils ETC

Les critères d'évaluation de l'outil ETC seront différents des outils de présentation, car l'outil de présentation est destiné aux utilisateurs finaux tandis que l'outil ETC sera dédié aux experts du service de l'informatique. Voici les points à considérer lors de l'évaluation :

a) **Interopérabilité** - Le système doit permettre l'accès rapide aux données stockées dans les bases de données Oracle puisque les données des systèmes transactionnels de l'UQTR sont emmagasinées dans une base de données Oracle. Le système devra aussi être en mesure de traiter les fichiers plats, car des données très intéressantes à des fins d'analyse (ex. : ICOPE) peuvent provenir de fichiers plats.

b) **Plan de déroulement de tâches**[4] – Le système doit offrir une interface pour la conception et l'exécution du déroulement des opérations de transformation nommées « *workflow* » nécessaires, au chargement des données dans l'entrepôt.

c) **Tâches et transformations** – Le système devra fournir une librairie de tâches et de transformations, sous la forme de composantes facilement réutilisables, qui permet la création de *workflow*. Typiquement, ces tâches et ces transformations permettent la copie, la suppression, la modification, la jointure, les calculs, la «dénormalisation», le triage, le filtrage, l'échantillonnage, etc. sur des flux de données (ex. : tables de BD, fichiers plats, fichiers XML, lien FTP, etc.).

[4] Ce terme est synonyme avec le terme « workflow » qui est typiquement utilisé en anglais.

d) **Séquenceur** – Le système doit permettre de spécifier des plages de temps pour l'exécution des plans de déroulement de tâches.

e) **Validation du chargement** – Le système doit permettre des mécanismes qui permettent de valider l'exécution d'un plan de déroulement de tâches.

f) **Gestion des erreurs** – Le système doit permettre des mécanismes de gestion d'erreurs (p.ex. : boucles et opérateurs conditionnels), idéalement tous les événements d'erreurs (non-critiques) devraient être stockés en tant que métadonnées et consultables lors de l'analyse des plans de déroulement des tâches. Ces résultats sont indispensables afin de permettre la gestion des données erronées (ex. : doublons, valeurs manquantes, valeurs inconsistantes et valeurs isolées).

g) **Endroit de staging 2** – Le système devrait permettre des endroits de stockage intermédiaires (« staging ») entre chaque étape importante du plan de déroulement de tâches telles que l'étape d'extraction, de nettoyage de données et de conformité des tables de dimensions et de faits. Ceci rend facile l'analyse et le diagnostic d'erreur lors de l'exécution du processus de chargement vers l'entrepôt. D'autant plus que cela évite d'exécuter à nouveau les étapes du processus antérieur lorsqu'il y a interruption.

Pour la synthèse des critères d'évaluation, cinq regroupements ont été créés. Chaque point sera détaillé en sous points dans la synthèse s'il y a lieu. Voici les cinq divisons : «nettoyage des données», «chargement des données», «transformation des données», «gestion des modèles dimensionnels» et «mise à jour de l'entrepôt».

5.2.2 Critères des logiciels de présentation des données

Les outils de présentation de données vont permettre aux utilisateurs finaux de créer leurs propres tableaux de bord personnalisés. Certains utilisateurs voudront seulement exécuter des rapports prédéfinis, d'autres voudront contrôler chaque étape de réalisation du tableau de bord. Il est aisé de constater qu'il y aura différents niveaux d'utilisateurs pour différents besoins analytiques. Le logiciel de présentation doit être simple d'utilisation et intuitif.

Les critères d'évaluation pour les outils de présentation sont classés en deux points : «les rapports et tableaux de bord» et «outils d'analyse».

Des critères qualitatifs ont été ajoutés à la synthèse. Ces critères sont seulement à titre indicatif et se basent sur les échanges qui ont été faites lors de l'évaluation. Ces critères sont : la «Qualité de la présentation des produits par les représentants» et le «Service à la clientèle (avant la vente)».

5.2.3 Évaluation détaillée des outils retenus

Cette évaluation est basée sur l'ensemble des produits d'entrepôt de données (DW/BI) qui ont été évalués à ce jour. Une enquête préliminaire a été réalisée sur plusieurs produits de différents fournisseurs tels que DMExpress, Sunopsis, Pentaho, Talend, IBM Red Brick[5], Clover.ETC, ContourComponents, Warehouse Builder d'Oracle 10G, JasperSuite de JasperSoft, PowerCenter d'Informatica, Crystal Decisions de Business Objects (ci-après nommé BO dans le texte), Data Manager de Cognos, Cognos8 de Cognos et finalement SAS. Il faut mentionner que par faute de temps, les outils BI de Microsoft ont été exclus de l'enquête. Suite à cette enquête préliminaire, certains produits ont retenu notre attention pour une évaluation plus approfondie :

a) Warehouse Builder d'Oracle 10G
b) SAS
c) Data Manager et Cognos8 de Cognos
d) PowerCenter d'Informatica
e) Crystal Decisions de Business Objects
f) ***Open source :*** JasperSuite **et** Pentaho

Cette première évaluation fut élaborée suite à de nombreuses conversations téléphoniques et des démonstrations de fournisseurs. Par la suite, chacun des produits a été coté sur un ensemble de critères fonctionnels et non fonctionnels. Les critères fonctionnels comprennent le nettoyage, le chargement, la transformation des données, la gestion des modèles dimensionnels et la mise à jour de l'entrepôt. Les critères non fonctionnels consistent en un ensemble de critères comme par exemple : la convivialité des interfaces,

[5] Il est important de prendre note que nous avons exclu le fournisseur IBM de notre évaluation dû au fait que son coût d'achat est trop élevé (supérieur à 120 000 $).

la création des tâches par assistant (« *Wizards* »), la qualité de la présentation et du service offert par les représentants.

Les critères d'évaluation peuvent être catégorisés en deux principaux volets:

- le mécanisme ETC (Extraction, Transformation et Chargement)
- la composante BI (outils d'analyse OLAP, rapport ad hoc et tableaux de bord)

Parmi les produits retenus, à la fin de notre première évaluation, notre choix s'est orienté davantage sur deux fournisseurs, dû à leurs démonstrations convaincantes et leurs résultats obtenus dans notre grille d'évaluation. Le tableau 5.5 permet de comparer les avantages et les inconvénients de chacun des produits se distinguant dans l'évaluation :

Tableau 5.5
Comparatif des composantes de la première étude

	INFORMATICA PowerCenter(ETC) et Meta-Data Manager	**COGNOS** Data Manager(ETC) et Cognos BI 8 (Présentation)
Composante ETC :		
Gestion des transformations	Environnement très convivial et pratiquement sans programmation manuelle.	Peut nécessiter l'ajout de programmation manuelle pour réaliser certains traitements.
Analyse et nettoyage des données	Offre plusieurs options pour l'analyse et le nettoyage de données.	Aucune composante d'analyse. Ceci doit se faire avec de la programmation manuelle.
Composante BI :		
Gestion et création des rapports et tableaux de bord	**AUCUN**	Environnement extrêmement convivial (entièrement Web) Outils très puissants pour la création de tableaux de bord (jauges, cartes, etc.).

Il serait important de considérer attentivement les bienfaits d'une configuration «hybride» qui utiliserait à la fois la composante ETC (PowerCenter) d'Informatica et la composante BI de Cognos, mais pour cela il faut avoir du budget.

Après la réception des coûts des fabricants, l'entreprise « Informatica » qui n'offrait que la composante ETC fut écartée de notre proposition à cause de son coût trop élevé. Le prix d'achat est de $400 000 pour un logiciel qui ne fait que la partie ETC. Notre budget initial

est une enveloppe de $80 000 pour à la fois l'outil ETC et l'outil de présentation des données.

Le lecteur trouvera en annexe «G», la grille synthèse de l'évaluation de cinq produits. De la synthèse de cette première évaluation, aucun point concluant ne permet de faire un «meilleur choix». Pour chaque produit, on évalue par un «oui» ou par un «non» si l'outil possède la fonctionnalité ou le service mentionné. Presque toutes les fonctionnalités sont présentes dans la plupart des produits. Nous étions alors dans l'obligation de réviser nos choix. D'autres rencontres furent planifiées mais cette fois-ci entre les compagnies COGNOS, BO et SAS. Une deuxième évaluation plus détaillée était donc nécessaire.

Afin de faire une proposition d'achat sur un de ces produits, il nous restait à comparer chaque produit avec des critères plus spécifiques. Il fallait aussi comparer les produits avec un même processus d'affaires, à l'implanter avec les mêmes données et préparer les mêmes tableaux de bord à la sortie. Les spécifications d'un processus d'affaires que l'on veut implanter à l'UQTR ont été établies et ont été soumises à chaque produit. Ce processus fut le suivi de l'admission. Il fallait suivre à la trace une demande d'admission d'un étudiant à un programme, en passant par les admissions conditionnelles jusqu'à son inscription à une première session ou à l'abandon de sa demande.

Il a fallu installer, tester et évaluer les outils ETL et les outils de présentation de chaque solution commerciale. Cette deuxième évaluation fut plus détaillée, plus technique avec un aspect quantitatif. Plusieurs mois ont été nécessaires afin de compléter cette étude. Cela a permis néanmoins de démontrer les avantages et les inconvénients de chaque produit en nous présentant toutes les étapes de l'implantation du processus d'affaires. Il est alors plus facile de comparer et de faire un choix éclairé. Le tableau 5.6 nous indique l'ensemble des critères ayant été établis pour comparer les logiciels

Tableau 5.6

Critères de comparaison des outils

1) INTÉGRATION DES DONNÉES (ETL) :
- a. Support "natif" pour Oracle CDC
- b. Utilisation de métadonnées (p.ex.cadre dimensionnel)*
- c. Acquisition de données (Extraction)
- d. Transformation des données (Transformation)
- e. Chargement des données (Load)
- f. Création et gestion des flots (p.ex. parallélisme)*
- g. Gestion des dimensions lentes (SCD) et des hiérarchies déséquilibrées*
- h. Gestion des changements de structures
- i. Outils à base d'assistant ("wizard")*
- j. Intégration avec gestion des métadonnées (voir section 2)
- k. Permet l'exécution par séquenceur ou l'invite de commande (shell)
- l. Le langage script et librairie de fonctions

2) GESTION DES MÉTA-DONNÉES :
- a. Gestion des tables ("query subject" et table dérivée)
- b. Utilisation d'alias de tables
- c. Définition des hiérarchies (niveaux et membres)*
- d. Gestion des boucles, "fan traps" et "chasm traps"
- e. Gestion de l'intégrité des modèles (univers, package, etc.)
- f. Outils de support à la conception
- g. Gestion de la sécurité et le contrôle des accès
- h. Convivialité de l'interface
- i. Analyse des impacts et "Data Lineage"

3) GESTION DES RAPPORTS :
- a. Spécification des champs de table (query item)
- b. Spécification des champs calculés (variable ou data item)
- c. Spécification du triage de champ
- d. Ajout d'un filtre
- e. Création des sections
- f. Spécification d'un group
- g. Actualisation des données
- h. Visualisation et interactivité avec la structure du rapport
- i. Possibilité de comparer diverses versions d'un rapport
- j. Convivialité des objets (constructs) et fonctionnalités en général
- k. Gestion des éléments de requêtes durant la création et gestion d'un rapport
- l. Ajustement et peaufinage des rapports actualisés
- m. Rapidité et temps réponse des écrans

4) ANALYSES ET EXPLORATIONS :
- a. Création de la structure principale du rapport d'analyse (tableau croisé)
- b. Utilisation des hiérarchies (niveaux et membres)
- c. Emplacement des niveaux ("Nesting",."Stacking")
- d. Triage
- e. Filtrage
- f. Ajouter des calculs (p.ex.: sous-totaux)
- g. Forage ("Drill", "Roll-up", "Slice", "Swap")
- h. Permet l'utilisation des cubes MOLAP ou outils d'analyse OLAP qui est "aggregate aware"
- i. Permet les requêtes MDX (Multi Dimensionnal Extension)
- j. Convivialité de l'interface
- k. Exploitation (données et forage) en utilisant Microsoft Office

5) INDICATEURS DE PERFORMANCE :
- a. Gestion des évènements
- b. Création des tableaux de bord (Dashboard)
- c. Création des cartes de pointage (Scorecard)

6) DÉPLOIEMENT :
- a. Plateforme et système d'exploitation
- b. Permet d'utiliser des APIs de programmation (p.ex.: Java) pour faire de l'intégration sur mesure à nos systèmes.

7) SOUTIENS LOCAL et de la COMMUNAUTÉ :
- a. Représentants sur place (Montréal et environs)*
- b. Accès à une communauté/base d'utilisateurs
- c. Qualité des présentations des produits par les représentants
- d. Service à la clientèle (avant la vente)

Les compagnies COGNOS, BO et SAS nous ont permis de tester leur suite. L'outil ETL et l'outil de présentation des données COGNOS furent installés à l'UQTR. La suite de BO fut testée à l'UQAM (Université du Québec à Montréal) entièrement par mon collègue Michel Charest avec la collaboration du Bureau de la Recherche Institutionnelle (BRI), sur un portable qu'ils nous ont prêté sur place tandis que la suite de SAS fut testée en collaboration avec la Direction de la recherche institutionnelle (DRI) au siège sociale de l' Université du Québec (UQ).

Les tests avec les suites de COGNOS et BO furent concluants tandis qu'avec SAS, l'installation sur notre portable par l'UQ n'a pu être complétée faute d'erreur logicielle. La compagnie SAS n'a pas voulu intervenir directement sans frais pour corriger le problème et nous permettre de tester en profondeur leur suite de logiciel puisque selon eux, leur prix d'achat très bas (moins de $25 000 négocié par l'UQ pour l'ensemble des universités du réseau) ne justifiait pas cette gratuité. Il est donc assez difficile de mettre un pointage équitable à ce produit que l'on a pu tester nous même en entier.

Mon collègue Michel Charest s'est vu mandaté pour l'évaluation technique de la grille finale des produits COGNOS et BO. De mon point de vue, il est évident que la même personne devait faire tout le processus sur les deux produits afin de bien quantifier les nuances. La marge est très mince entre un pointage 1,2,3 ou 4 pour un critère donné.

Dans un premier temps, chaque outil a été implanté et testé avec le même processus d'affaires et ce dans les deux produits. Une grille découpée en 10 points indique ce qu'il faut tester. Par exemple la définition et la gestion des sources et des connexions ou encore la gestion des dimensions. Les tableaux 5.7 à 5.9 expliquent chaque point et sous points de la grille.

En plus des produits commerciaux COGNOS, BO et SAS, vous trouverez en annexe «H» la comparaison des logiciels libres (*Open source*) JasperSoft et Pentaho basée sur les mêmes critères d'évaluation que ceux du tableau 5.6.

Cette évaluation nous a permis d'affiner nos besoins en terme de nombre d'utilisateur, de plate-forme et d'autres points techniques plus fins tel que l'utilisation d'un séquenceur, la

possibilité de forage entre différents rapports directement intégrés dans l'outil de présentation des données et plusieurs autres. Vous trouverez à l'annexe «I» le devis pour l'achat d'un système d'intelligence d'affaires résumé par Michel Charest. Dans la majeure partie des points énoncés dans cette annexe, c'est le produit BO qui se démarque sur toute la ligne

Tableau 5.7
Plan technique de comparaison des outils (points 1 à 3).

#	ESSAIS ou CRITÈRE D'ÉVALUATION	EXPLICATION
1)	**Activité 0 - Importation des métadonnées :**	
.1	Définition et gestion des sources et connexions	Est-il facile de gérer les connexions aux sources OLTP ?
.1.1	Compatibilité avec Oracle 10g	Est-ce que les connexions aux sources Oracle sont relativement stables et efficaces ?
.2	Définition et gestion des sujets d'affaires	Est-il facile de définir des sujets d'affaires (tables et colonnes) ?
.3	Niveau d'importation permis	Est-il possible d'importer des éléments OLTP à un niveau fin (tables, vues, colonnes, fonctions, etc.) ?
.4	Importation des sujets d'affaires et création de modèles	Est-il relativement facile de créer des modèles (p.ex.: schéma étoiles, flocons) ?
.4.1	Gestion des clés étrangères	Est-ce que l'application permet de facilement définir les relations entre les sujets d'affaires ?
.5	Gestion des dimensions	Est-ce que l'application permet de facilement définir les clés et attributs pour les tables de dimension ?
.5.1	Définition et gestion des hiérarchies	Est-ce que l'application permet de facilement définir les hiérarchies et niveaux? Par exemple : DÉPARTEMENTS -> CYCLES -> PROGRAMMES
.6	Gestion des tables de faits	Facile de spécifier les faits et attributs (et leurs types de données correspondantes) ?
.6.1	Gestion des mesures	Est-il facile de définir et gérer les types d'agrégations à apporter aux mesures (additive, semi-additives, etc.) ?
.7	Gestion des vues d'affaires	Existe-t-il des mécanismes pour offrir diverses vues de présentation aux utilisateurs (Interne, Affaire et Dimensionnelle) ?
.8	Support le Common Warehouse Meta-Model (CWM)	Est-ce que l'application permet de stocker le modèle d'entrepôt sous le format CWM ? Ceci permet une portabilité entre les divers produits BI
.9	Publication des métadonnées et structure vers le portail web	Habituellement une application permet de préparer un paquet bien défini qui sera publié dans l'environnement utilisateur.
.10	Mécanisme de vérification des objets	Avant de publier une "paquet" sur le portail, est-ce que celui-ci est validé pour assurer son bon fonctionnement ?

.11	Gestion des versions de "paquets" publiés aux usagers	Est-ce que l'outil permet de stocker les versions successives d'un paquet ? (il devrait être possible de retourner à une version antérieure d'un paquet)
2)	**Activité 1 - Création d'un rapport simple :**	
.1	Gestion du type de rapports	Est-ce que l'outil permet l'utilisation de formats de rapports prédéfinis (p.ex.: liste, tableau-croisé, figure, groupe à répétition, graphique etc.) ?
.2	Emplacement des objets	Il devrait être relativement facile d'ajouter, modifier et de supprimer les éléments d'un rapport (p.ex.: texte, tableaux, marges, colonnes, titres, etc.)
.3	Gestion de la structure de la page	Est-ce que l'application permet de visionner un rapport sous plusieurs vues (p.ex.: conception, structure, entête et bas de pages, actualisation avec données) ?
.4	Intégration des données à un rapport	Est-il relativement facile d'ajouter des sujets d'affaires dans un rapport (colonne, niveau ou membre hiérarchique) ?
.5	Format de rapports supportés	Est-il facile de produire les rapports en différents formats (PDF, HTML, Excel, CSV, XML, etc.)?
.6	Application d'opérateurs de base :	
.6.1	Regroupement des données (GROUP BY)	S'agit-il simplement de sélectionner les colonnes et un bouton ?
.6.2	Tri des données (ORDER BY)	S'agit-il simplement de sélectionner les colonnes et un bouton ?
.6.3	Filtrage de données (WHERE x ...)	S'agit-il simplement de sélectionner les colonnes et un bouton ?
3)	**Activité 2 - Création d'un rapport avancé :**	
.1	Insertion de plusieurs objets graphiques sur une page	Est-il facile d'insérer et de modifier les objets graphiques sur le rapport ?
.2	Réalisation d'une carte «forable»	Est-il possible de réaliser une carte graphique avec des zones «forables» (pour plus de détails) ?
.3	Besoins spécialisés / requêtes sur mesure	Est-il possible d'utiliser des requêtes SQL ou MDX pour des besoins de rapports plus avancés et particuliers (autre que les sujets d'affaires) ?
	Lien forage d'un rapport à un autre	Il est facile de configurer des liens «forables» entre rapports ?
.4	Utilisation de variables pour le formatage conditionnel	Est-il possible spécifier des règles pour la présentation d'un rapport (p.ex. (if vente > 10 000 then 'cell=rouge')

Tableau 5.8
Plan technique de comparaison des outils (points 4 à 6).

#	ESSAIS ou CRITÈRE D'ÉVALUATION	EXPLICATION
4)	**Activité 3- Une analyse OLAP :**	
.1	Création d'un tableau croisé	Facile d'insérer les dimensions et faits afin de réaliser un tableau «forable» ?
.2	Opération de bases	Facile d'effectuer des opérations supplémentaires sur les données tel que le triage, formatage, sommes (calculs), sélection/suppression de colonnes, appliquer des filtres, etc. ?
.3	Opération de forage	Possible de facilement faire du *drill-down, drill-up*, slice (cibler un membre particulier), *swappage* des axes, etc. ?
.4	Outils supplémentaires pour faciliter l'analyse des données	Possible de créer des graphiques supplémentaires (ligne, histogramme, etc.) pour faciliter l'analyse ?
.5	Format de stockage des analyses	Possible de sauvegarder une analyse en format HTML, PDF, CSV, XML, etc.) ?
5)	**Activité 4 - Création d'un tableau de bord :**	
.1	Ajout de jauges (pour indicateur de performance)	Facile à configurer ?
.2	Ajout d'un tableau croisé «forable»	Facile à configurer ?
.3	Ajout d'objet graphique	Facile d'ajouter et configurer des graphiques du genre histogramme, pointe de tarte, etc. ?
.4	Support adéquat pour les objets graphiques	Possible de créer une gamme de différents graphiques de genre ligne, histogramme, pointe de tarte, polaire, etc.) ?
6)	**Flexibilité pour les utilisateurs (aspect qualitatif)**	
.1	Le stockage des "vues" de rapport	Possible de stocker une configuration de rapport en entier avec ses paramètres d'exécution ?
.2	Intégration de l'ensemble des outils BI	L'environnement devrait posséder un nombre suffisant d'outils (mais pas trop). Est-il facile d'interagir entre les diverses composantes ?
.3	Uniformités des fonctionnalités	D'un composante à l'autre, est-ce que les fonctionnalités générales sont similaires (glisser/coller, barre d'outils, etc.) ?
.4	Qualité de la documentation	Est-ce que le niveau d'organisation et de détails est suffisant pour permettre un apprentissage ou dépannage rapide par l'utilisateur ?

Tableau 5.9
Plan technique de comparaison des outils (points 7 à 10).

#	ESSAIS ou CRITÈRE D'ÉVALUATION	EXPLICATION
7)	**Flexibilité pour les analystes/concepteurs (aspect qualitatif)**	
.1	Besoins spécialisés / programmation	Pour les tâches plus complexes, est-il possible d'intégrer de la programmation (p.ex.: PL/SQL, Java, etc.) pour combler un besoin particulier?
.2	Intégration de l'ensemble des outils BI	voir 6.2
.3	Uniformité des fonctionnalités	voir 6.3
.4	Qualité de la documentation	voir 6.4
8)	**Extractions, transformations et chargements :**	
.1	Définition et exécution des flots de travail	Environnement graphique facile à utiliser ?
.2	Utilisation des opérateurs de transformations	Propriétés des opérateurs et bloques de traitements dans les flots sont faciles à définir et utiliser ?
.3	Gestion des erreurs de traitement	Facile à détecter et gérer les cas d'exception dans les flots de travail ?
.4	Mécanisme de trace d'origine des données ("*Data Lineage*")	Possible de repérer le lien entre la donnée cible et la donnée d'origine source (souvent par l'entremise de métadonnées gérées par le système) ?
.5	Analyse des impacts par flots ETL ("*Impact Analysis*")	Possible que l'outil explique les impacts de modifier un flot de travail ETL sur les "paquets" publiés aux usagers ?
.6	Tableau de bord sur les états d'opérations ETL	L'outil offre une vue globale (forme de tableau de bord) sur l'état même des opérations des traitements ETL de l'entrepôt ?
9)	**Gestion des métriques / indicateurs de performance :**	
.1	Définition des flots stratégiques	Possible de spécifier des conditions mesurables qui appuieront la réalisation d'objectifs stratégiques
.2	Gestion des métriques de performance (indicateurs)	Possible de spécifier des fourchettes et d'y assigner un état en utilisant des couleurs (fort, moyen, faible) ?
10)	**Création des évènements BI:**	
.1	Création et diffusion d'événements	Possible de spécifier des conditions ou règles qui permettent d'informer rapidement les usagers de situations importantes (p.ex.: envoyer un courriel ou message sur portail si le taux d'admission baisse de 20% !)

Pour chaque point, l'échelle fut la suivante : Inconnu (?), Non applicable (!), Inacceptable (0), Faible (1), Bien (2), Excellent (3). Cette évaluation a permis pour l'outil COGNOS d'obtenir une cote de 130 tandis que l'outil BO s'est gratifié d'une cote de 141. L'écart n'étant pas très significatif il fallait continuer l'évaluation. Nous n'avons pu suivre ce plan avec la suite de SAS.

Tableau 5.10
Synthèse détaillée de la comparaison des outils.

#	Poids	Cognos Pts.	Cognos Total	BO Pts.	BO Total	SAS Pts.	SAS Total	Aperçu Qualitatif
1)								
a.	2	3	6	3	6	2	4	BO
b.	2	4	8	3	6	2	4	Cognos
c.	2	3	6	3	6	3	6	comparable
d.	3	2	6	3	9	1	3	BO
e.	2	3	6	3	6	3	6	comparable
f.	2	4	8	3	6	2	4	Cognos
g.	3	4	12	3	9	3	9	Cognos
h.	1	1	1	1	1		0	comparable
i.	2	3	6	1	2	1	2	Cognos
j.	1	3	3	3	3	1	1	comparable
k.	2	3	6	3	6	1	2	comparable
l.	3	3	9	3	9	2	6	comparable
				77		69		47
2)								
a.	1	3	3	3	3		0	comparable
b.	3	2	6	3	9		0	BO
c.	3	2	6	3	9		0	BO
d.	2	2	4	3	6		0	BO
e.	1	2	2	3	3		0	BO
f.	2	2	4	3	6		0	BO
g.	2	2	4	3	6	3	6	BO
h.	2	2	4	4	8	2	4	BO
i.	2	1	2	3	6		0	BO
				35		56		10
3)								
a.	1	3	3	3	3	2	2	comparable
b.	1	2	2	3	3	2	2	comparable
c.	1	3	3	3	3	2	2	comparable
d.	1	3	3	4	4	3	3	BO
e.	1	3	3	3	3	3	3	comparable
f.	1	3	3	3	3	3	3	comparable
g.	2	3	6	4	8	2	4	BO
h.	2	2	4	3	6	2	4	BO
i.	2	1	2	4	8	3	6	BO
j.	3	2	6	3	9	1	3	BO
k.	2	3	6	3	6	1	2	comparable
l.	2	3	6	2	4	1	2	Cognos
m.	3	1	3	4	12		0	BO
				50		72		36

#	Poids	Cognos Pts.	Cognos Total	BO Pts.	BO Total	SAS Pts.	SAS Total	Aperçu Qualitatif
4)								
a.	2	3	6	4	8	2	4	BO
b.	2	3	6	3	6	1	2	comparable
c.	2	3	6	3	6	1	2	comparable
d.	2	3	6	3	6	2	4	comparable
e.	2	3	6	3	6	2	4	comparable
f.	2	3	6	3	6	1	2	comparable
g.	2	3	6	3	6	2	4	comparable
h.	3	3	9	3	9	2	6	comparable
i.	1	1	1	3	3		0	BO
j.	3	2	6	4	12	2	6	BO
k.	2	3		3		2	4	comparable
				58		68		38
5)								
a.	2	3	6	2	4	1	2	Cognos
b.	3	3	9	3	9	2	6	comparable
c.	2	3	6	2	4	1	2	Cognos
				21		17		10
6)								
a.	2	3	6	3	6	2	4	(dépendra de nous)
b.	2	3	6	3	6	1	2	comparable
				12		12		6
7)								
a.	2	3	6	3	6	3	6	Cognos
b.	2	3	6	3	6	2	4	BO
c.	2	3	6	1	2	2	4	Cognos
d.	2	3	6	0	0	0	0	Cognos
				24		14		14
	107		277		308		161	BO
			64%		74%		38%	

Après que chaque outil fut testé avec le plan technique des tableaux 5.7 à 5.9, le tableau 5.10, représentant l'évaluation des outils en fonction de la grille d'évaluation du tableau 5.6, fut complété. Chaque point est quantifié pour chacun des outils COGNOS, BO et SAS. L'évaluation de SAS est basée sur les présentations que l'on a vues, sur quelques fonctionnalités testées et non en fonction du plan technique intégral. Afin d'être lisible, seul les pointages et les numéros des critères sont présentés dans ce tableau.

Dans le tableau 5.10, le poids représente l'importance du point en fonction de nos besoins. La colonne «pts.» indique le pointage de la compagnie pour le critère. La colonne «total»

représente la multiplication du poids par le pointage de la compagnie. Le nombre sur la dernière ligne d'une section représente la somme des pointages pour la compagnie et finalement la colonne «Aperçu qualitatif» indique la compagnie qui ne se démarque pas nécessairement au pointage.

5.3 Recommandations

Voici en résumé, tout le processus d'évaluation avec les conclusions et les recommandations.

Enquête préliminaire

- **Date :** Août 2007 à octobre 2007
- **Description :** Inventaire des outils
- **Conclusion :** Choix de certains outils pour une étude plus approfondie
- **Liste des outils (ceux en caractères gras ont été retenus pour l'évaluation #1)**
  - DMExpress
  - IBM Red Brick
  - Warehouse Builder d'Oracle 10G
  - Sunopsis
  - Clover.ETC
  - **JasperSuite de JasperSoft**
  - **Pentaho**
  - ContourComponents
  - **PowerCenter (Informatica)**
  - Talend
  - **Data Manager (Cognos)**
  - **Crystal Decisions de Business Objects (BO)**
  - **SAS**
  - **Cognos8 (Cognos)**

Évaluation #1 (annexe «G» pour la synthèse de l'évaluation)

- **Date :** Novembre et décembre 2007
- **Description :** Études des outils sélectionnés à l'enquête préliminaire
- **Liste des outils évalués :**
  - **Pentaho**
  - **JasperSuite**
  - **PowerCenter (Informatica)**
  - **SAS**
  - **Cognos (Data Manager et Cognos 8)**
  - **Crystal Decisions de Business Objects (BO)**
- **Conclusion :** Dans un premier temps la solution «hybride» des compagnies Cognos et Informatica fut sélectionnée. Par son prix trop élevé ($400 000), Informatica qui offre seulement l'outil ETC fut exclus. Il n'y a alors aucune autre solution retenue.

- **Recommandation :** Faire une deuxième évaluation plus détaillée et plus technique en excluant l'outil PowerCenter d'Informatica.

Plan technique des étapes à réaliser avec les logiciels

- **Date :** Décembre 2007 à février 2007
- **Description :** Réalisation de la grille technique des tableaux 5.7 à 5.9 et études des outils COGNOS et BO avec cette grille.
- **Conclusion :** La compagnie COGNOS a obtenu une cote totale de 130 tandis que l'outil BO s'est gratifié d'une cote de 141. La suite BO est légèrement supérieure. Nous n'avons pu évaluer SAS avec cette grille.
- **Recommandation :** Poursuivre l'évaluation en complétant les critères du tableau 5.6.

Évaluation #2 (annexe «H» pour les outils *Open sources*)

- **Date :** Février 2007 à mars 2007
- **Description :** Évaluation détaillée des outils en fonction des critères du tableau 5.6 et suite aux tests réalisés avec le plan technique.
- **Conclusion :** Le tableau 5.10 résume en détail pour les compagnies COGNOS, BO et SAS les pointages respectifs de 64%, 74% et 38%. Le résultat de la suite BO est supérieur. Le tableau 5.11 résume sommairement le pointage en plus avec les deux compagnies Jasper et Pentaho.
- **Recommandation :** La suite BO est recommandée pour l'achat.

Devis pour la recommandation d'achat (annexe «I»)

- **Date :** mai 2007
- **Description :** Devis des besoins techniques des outils ETC/BI pour l'entrepôt.
- **Recommandation :** La suite BO est recommandée pour l'achat car elle répond à ces autres besoins complémentaires.

Le tableau 5.11 nous montre le pointage final de chaque grande division de l'évaluation de la grille des critères du tableau 5.5. Nous recommandons l'achat de la suite BO. Un des avantages de cette suite est l'utilisation du langage de programmation PL/SQL qui est le langage utilisé à l'UQTR et surtout par la simplicité de ses interfaces pour les utilisateurs

dans l'outil de présentation des données. Par contre, l'outil COGNOS est hautement flexible ce qui apporte une complexité certaine mais avec plus de possibilités.

Tableau 5.11
Synthèse finale de la comparaison des outils.

	COGNOS	BO	SAS	Jasper	Pentaho
1) Intégration des données	77	69	47	46	51
2) Gestion des métadonnées	35	56	10	18	18
3) Gestion des rapports	50	72	36	47	2
4) Analyses et explorations	58	68	38	43	43
5) Indicateurs de performance	21	17	10	5	0
6) Déploiement	12	12	6	12	12
7) Soutient	24	14	14	4	6
Total :	277	308	161	175	132
	64%	74%	38%	42%	32%

La recommandation de la suite de Business Object (BO) s'impose. Elle combine les résultats du plan technique d'évaluation (tableaux 5.7 à 5.9), de l'évaluation détaillée des outils (tableaux 5.10 et 5.11) utilisant les critères du tableau 5.6 et finalement des besoins du devis de l'annexe «I».

Le résultat de l'évaluation des outils n'aurais pas été le même si elle avait été faite deux ans plutôt. Comme on ne connaît pas l'avenir, ce résultat est le meilleur au moment d'écrire ces lignes. Il ne faut pas oublié que la suite de Microsoft n'a pas été évaluée et que celle de SAS n'a pu être complétée. Actuellement, le développement logiciel est très prolifique dans ce domaine. Il y a beaucoup de fluctuations entre les compagnies : vente et achat de compagnie par une autre, fusion, partenariat. Ces événements pourront influencer les leaders de demain. Cette analyse résulte donc des outils considérés pendant l'étude en cours durant l'année 2007.

La demande d'achat de la suite BO est présentement en cours. Nous prévoyons l'installation de la suite à la session d'automne 2008. À la session d'hiver 2009, nous débuterons la création d'un premier processus d'affaires afin de produire notre premier tableau de bord pour la fin de la session d'hiver 2009.

Partie 3
Conception de l'entrepôt de données

Dans cette partie, un prototype sera mis en place en suivant la méthodologie proposée au chapitre 4 tout en appliquant les phases de développement des processus ETC. Le processus de conception s'étend de la phase d'intégration à la phase de publication des données. Cette dernière permettra aux dirigeants de construire leurs tableaux de bord. Cette troisième partie est présentée en quatre chapitres, soient :

6. Conception du modèle
7. Publication des données
8. Résultats et travaux futurs
9. Conclusion

Chapitre 6
Conception du modèle

Ce chapitre nous guidera dans l'application de la méthodologie. À la fin de ce chapitre, les données seront intégrées au prototype de l'entrepôt de données. Quelques problèmes rencontrés ont nécessité la réalisation d'outils. Ces outils serviront à la fois au SSPT dans le développement logiciel et à l'analyse et la préparation des données pour la mise en place de l'entrepôt de données :

6.1 Préparation des données
6.2 Modélisation dimensionnelle
6.3 Modèle physique de l'entrepôt
6.4 Extraction, transformation et chargement de l'entrepôt

6.1 Préparation des données

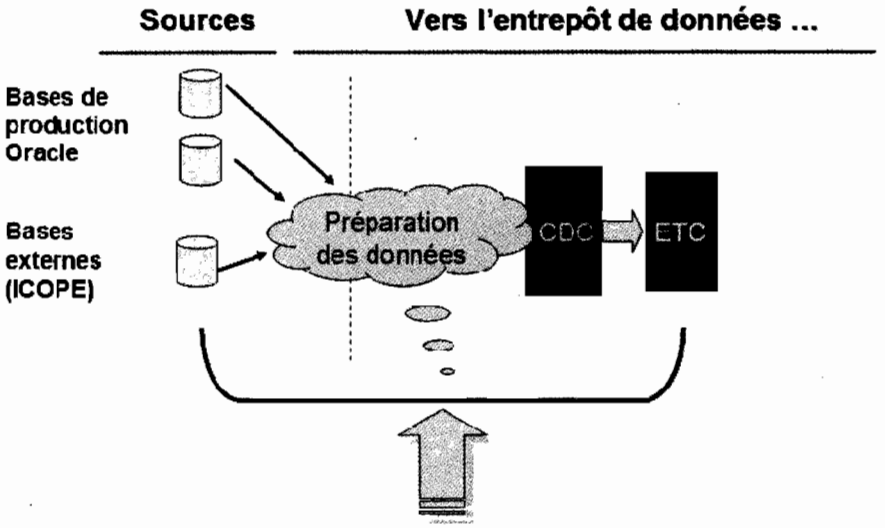


Figure 6.1 Zone de préparation des données.

Un ETC permet de charger deux types de données : les données agrégées ou les données unitaires (transactionnelles).

Ces données chargées, provenant des systèmes transactionnels, sont considérées comme les données sources qui seront intégrées à l'entrepôt pour devenir les données cibles. Il sera important, lors de la définition du processus d'affaires, que l'analyste informatique qui le réalisera possède une profonde connaissance des structures du SGBD associées aux données qui seront chargées.

Dans l'étude du chapitre 1, les analystes responsables de chaque système de gestion de l'UQTR sont identifiés. Lors de l'élaboration d'un processus d'affaires les concernant, ils seront invités à participer afin d'en maximiser et d'en optimiser tout le processus.

Le repositionnement des phases de la méthodologie effectuées aux chapitres précédents est redéfini comme suit :

1. approche de base «bottom-up»;
2. orientée «objectifs et besoins»;
3. dimension et fait conformes;
4. **cycle de vie dimensionnel de Kimball**
    a. avec arbre du sujet pour la définition des dimensions;
5. moteur ROLAP.

Les points 1,2 et 3 orientent le cycle de vie décisionnel. Dans ce chapitre, ne seront pas abordé les plans de «définition de l'architecture technique» ni celui de «l'installation et la sélection des produits». Le moteur ROLAP est plutôt en fonction de l'outil de présentation qui sera choisi et installé. L'étape quatre est celle qui est à réaliser. Elle a été bonifiée de la schématisation de l'arbre du sujet. La figure 3.16 (page 43) nous sert de base à la mise en œuvre du cycle de vie.

La préparation des données mettra sur papier les objectifs généraux du service approché, les besoins en analyse et la disponibilité des données dans les OLTP.

Monsieur Rémy Auclair, agent de recherche à la direction des affaires départementales, a répondu au questionnaire sur l'activité et les objectifs des utilisateurs. Vous trouverez aux tableaux 6.1 et 6.2 un résumé des réponses.

Tableau 6.1
Définir les objectifs généraux de la direction des affaires départementales

Définition du contexte général
Relevant du Vice-recteur aux études de premier cycle et au soutien académique et à la Vice-rectrice aux études de cycles supérieurs et à la recherche, la Direction des affaires départementales assume des responsabilités dans trois secteurs d'activités : les affaires départementales, la gestion des activités d'enseignement et la recherche institutionnelle.
Définition des objectifs (secteur : recherche institutionnelle)
La DAD produit et rend disponible diverses données portant notamment sur l'effectif étudiant, les postes de professeurs et les activités d'enseignement. Elle procède périodiquement à une enquête auprès des diplômés de baccalauréat et de maîtrise, et participe à des projets touchant à la réussite des étudiants.
Définition des objectifs généraux
▪ Données descriptives sur l'effectif étudiant et les activités d'enseignement. ▪ Données sur les postes de professeurs, sur les tâches et les activités d'enseignement. ▪ Relance auprès des diplômés. ▪ Le suivi de l'admission. ▪ Le suivi de la clientèle sur 5 ans. ▪ Autres dossiers d'analyse.

Cette étape est ajoutée pour pallier à la faiblesse de l'approche «piloter par les besoins». Il convient de décrire les objectifs globaux du Service avant de poursuivre. Cette approche est dite «piloter par les objectifs». Par la suite, un processus d'affaires que l'on voudra intégrer à l'entrepôt de données devra être en lien avec les objectifs généraux du Service.

Le premier objectif choisi pour une première modélisation est «le suivi de l'admission».

Tableau 6.2
Définir les besoins de la direction des affaires départementales

Définition du processus	
Suivre les diverses étapes du processus d'admission à un trimestre donné.	
Description détaillée du besoin	
«Suivre progressivement les étudiants à partir de l'étape de la demande d'admission en passant par les candidats acceptés, les nouveaux inscrits (dans l'établissement et dans le programme) et les inscriptions totales. Le tout serait disponible selon un certain nombre de caractéristiques : le programme, le genre de programme, le sexe, le régime d'études, le groupe d'âge, le collège et la région de provenance, la CRC et le diplôme comme base d'admission.»	
Quelques définitions du domaine	
Demandes d'admission	Toute demande d'admission à un programme d'études universitaires.
Candidats acceptés	Toute personne ayant présenté une demande d'admission et ayant reçu une offre d'admission.
Nouveaux inscrits	Toute personne s'inscrivant pour la première fois dans un programme, **incluant** les changements d'orientation.
Inscriptions totales	Nombre total de personnes qui se sont inscrites à des programmes.
Mesures	
Nombre d'admissions dans un programme à une session. Nombre de candidats acceptés dans un programme à une session. Nombre de «nouveaux inscrits» dans un programme à une session.	
Calculs	
Somme, moyenne et pourcentage de chaque mesure.	

Tableau 6.2 (suite)
Définir les besoins du service

Données			
Donnée désirée	**Table source**	**Donnée désirée**	**Table source**
programme	adm_formulaire_pgm (cd_pgm)	genre de programme	tri_tb_secteur_pgm (desc_sect_pgm **ou** cd_classe_pgm)
sexe	gen_personne (sexe)	régime d'études	daf_tb_regi (regime_lng)
groupe d'âge	gen_personne (dt_naissance)	collège de provenance	daf_tb_inst (inst_lng)
cote R collégial	daf_dictionnaire (cote_rendement)	région de provenance	daf_tb_pgmc (pgm_colg_lng)
base d'admission	daf_tb_badm (desc_babmi)	cycle	tri_programme (cycle_pgm)
département	gen_tb_unit_sous_unti (desc_unit_sous_unit)	cd_perm	Identification de l'étudiant
Source des mesures			
Candidats acceptés	daf_etud_cand (cand_valid = 'O')		
Nouveaux inscrits	daf_etud_insc : présence d'une seule session (si session_insc = session_adm alors nouveaux inscrits)		
Critère de réussite			
➢ Avoir accès facilement et rapidement aux données ➢ Réduire le temps requis pour effectuer les analyses			

Une fois le questionnaire rempli, une rencontre avec l'analyste responsable du système du dossier étudiant a permis d'établir que les données requises par la demande étaient toutes existantes et présentes dans la même base de données. On poursuit donc la démarche du cycle de vie.

6.1.1 Analyse des données

Le demandeur sait que les données qu'il cible pour son analyse sont présentent dans les systèmes transactionnels. Le travail n'est pas terminé pour autant. Il ne suffit pas de voir uniquement le conteneur pour poursuivre, il faut aussi analyser le contenu.

La phase d'analyse des données sources doit être faite avant ou après la définition du schéma dimensionnel. Pour savoir quelles structures analyser, il faut schématiser le processus d'admission du système OLTP.

La modélisation en schéma ER pour le processus du suivi de l'admission est composée d'une douzaine de tables. Chaque table doit être analysée avant de débuter la modélisation dimensionnelle et la modélisation physique de l'admission dans l'entrepôt de données.

Regardez attentivement le schéma ER de la figure 6.2 de la page suivante. Il est très complexe et l'usager qui ne travaille pas avec ces structures aura beaucoup de difficulté à s'y retrouver. Imaginez la requête SQL sous-jacente!

Prenez note qu'une table n'est pas reliée aux autres. Il s'agit de la table qui représente la mesure «nouveaux inscrits». C'est par l'étape de transformation que l'on vérifiera si pour la même session que l'admission, le candidat est nouvellement inscrit.

La tâche d'analyse du contenu est très longue. Cependant, elle permet de corriger certaines erreurs à la source, ce qui évitera des erreurs au niveau des résultats, mais aussi l'application de modifications des données à l'entrepôt. Rendue à cette étape, une première problématique voyait le jour : le temps d'analyse. Le manque d'outils a fait perdre beaucoup de temps. De nombreuses requêtes SQL ont été créées manuellement. Deux types d'analyse ont été effectués. La première permettant de consulter les données d'une table, la seconde permettait de vérifier l'intégrité des données entre deux tables. Que de surprises nous attendaient! Une deuxième problématique s'annonçait. Sur certaines relations, aucune clé étrangère n'était définie. Il est alors devenu prioritaire de créer des outils afin d'accélérer le processus d'analyse en profondeur.

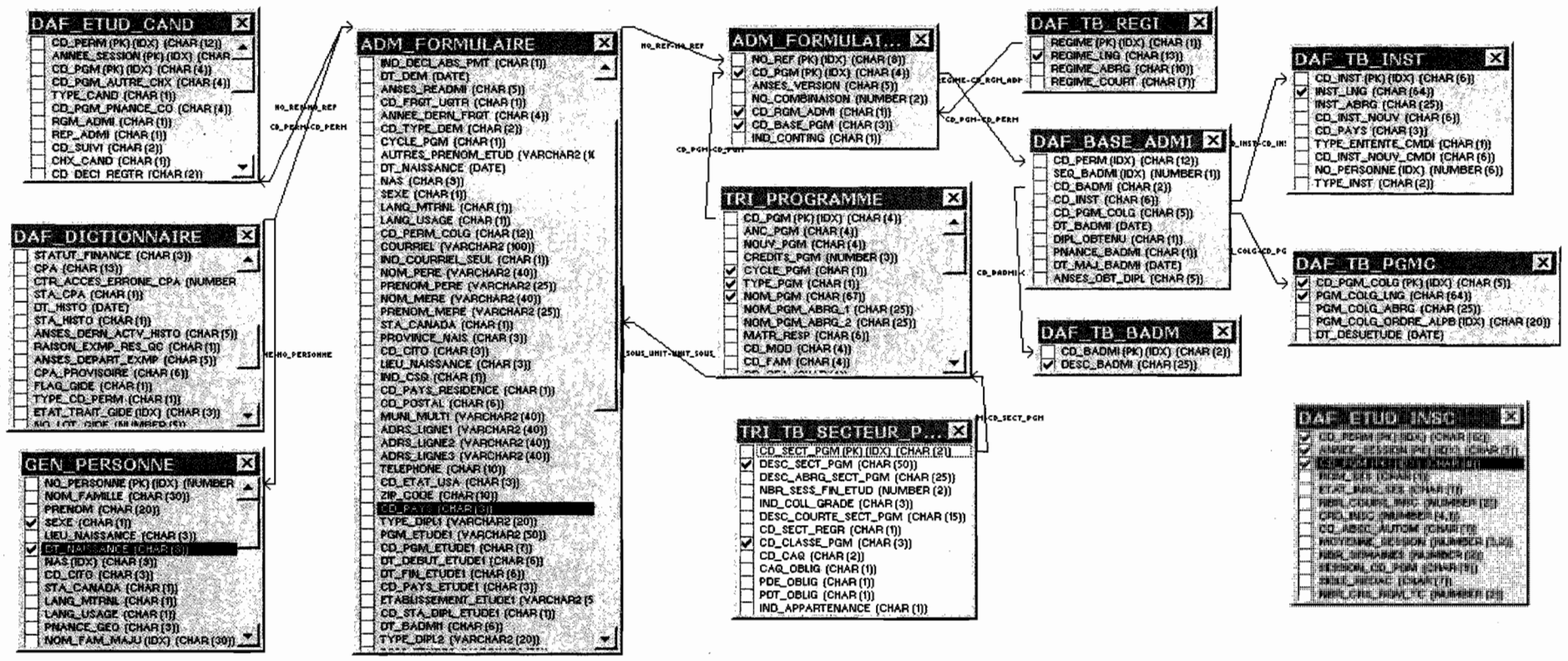


Figure 6.2 Schéma entité-relation du suivi de l'admission.

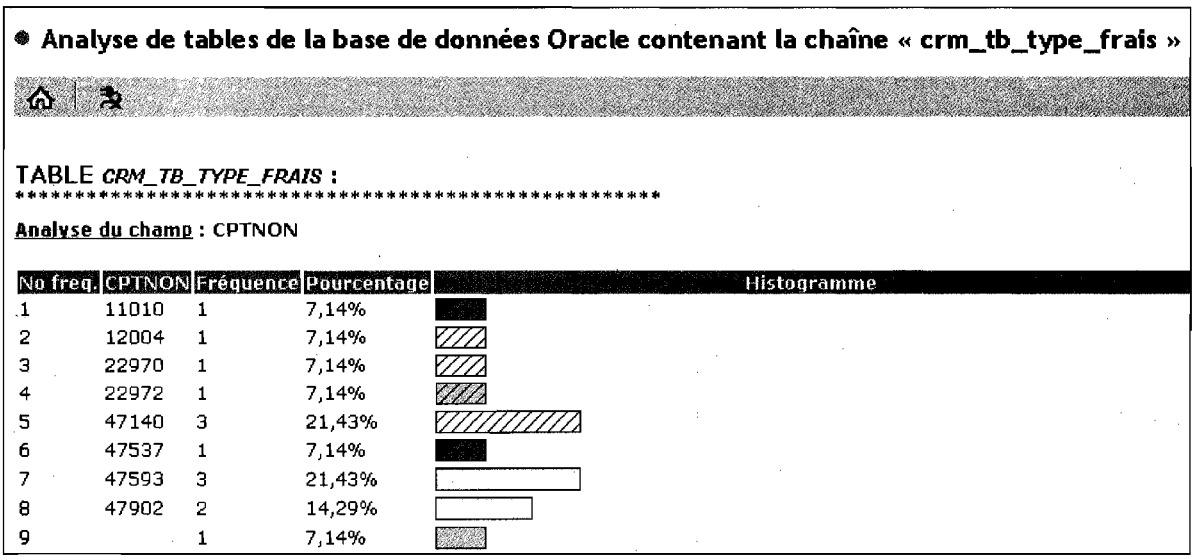
Une troisième problématique doit être soulevée. En pleine évaluation de la qualité des données des systèmes sources, un administrateur a appliqué des modifications de structures aux tables sur lesquelles les analyses étaient dirigées. Qui dit problématique, dit solution! Sans plus tarder, un autre outil de détection des modifications aux structures fut créé. La rapidité avec laquelle on réagit à une problématique de prétraitement est très bénéfique pour l'entreprise. Si on corrige à la source rapidement, on publie de bonnes données à long terme.

De cette dernière problématique en découle une quatrième. Si les structures des systèmes sources sont modifiées dans le temps, celles de l'entrepôt pourront l'être aussi. Il pourrait se produire un saut inexplicable dans les données résultantes pour l'usager.

6.1.2 Système d'analyse de tables (SAT)

Afin de pallier aux deux problématiques rencontrées, le système d'analyse de tables fut mis en place. Il permet l'analyse de fréquence des valeurs d'une table et l'analyse des relations (sur un seul champ ou sur plusieurs champ) afin de déterminer si une clé étrangère peut être appliquée entre deux tables si celles-ci on déjà des enregistrements.

L'outil offre les deux types d'analyse : analyse de la fréquence des valeurs et l'analyse de relations.



Analyse de tables de la base de données Oracle contenant la chaîne « crm_tb_type_frais »

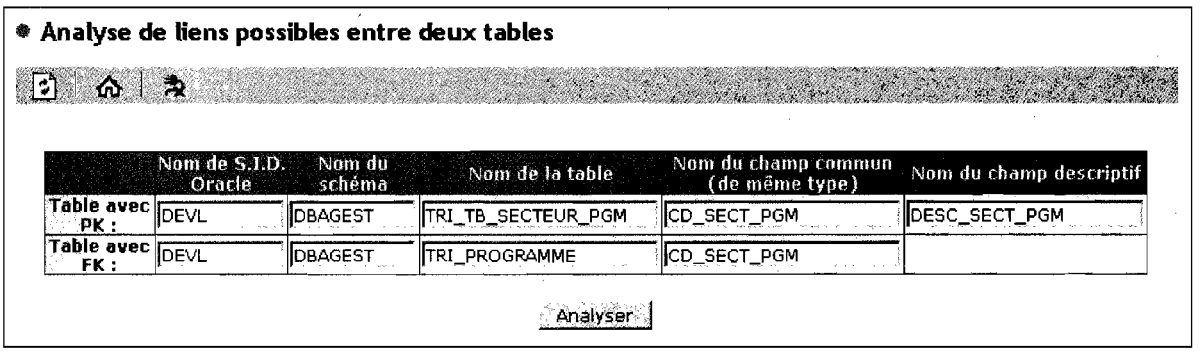
TABLE *CRM_TB_TYPE_FRAIS* :

Analyse du champ : CPTNON

No freq.	CPTNON	Fréquence	Pourcentage	Histogramme
1	11010	1	7,14%	
2	12004	1	7,14%	
3	22970	1	7,14%	
4	22972	1	7,14%	
5	47140	3	21,43%	
6	47537	1	7,14%	
7	47593	3	21,43%	
8	47902	2	14,29%	
9		1	7,14%	

Figure 6.3 Analyse de contenu avec SAT.

La procédure d'analyse de la fréquence cherche dans le dictionnaire Oracle, toutes les structures contenant la chaîne de texte tapée dans le but d'analyser la ou les tables dont le nom correspond à cette chaîne. On peut taper le nom d'une seule table, les premières lettres d'un système ou les dernières lettres avec le caractère générique «%». Pour chaque table résultante de la recherche, chaque champ est analysé. Dans la figure 6.3, le champ CPTNON de la table CRM_TB_TYPE_FRAIS contient neuf valeurs différentes, chacune utilisée à des fréquences différentes représentée en pourcentage et en histogramme. On peut voir les valeurs «null», les «blancs» et les données non utilisées.



Analyse de liens possibles entre deux tables

	Nom de S.I.D. Oracle	Nom du schéma	Nom de la table	Nom du champ commun (de même type)	Nom du champ descriptif
Table avec PK :	DEVL	DBAGEST	TRI_TB_SECTEUR_PGM	CD_SECT_PGM	DESC_SECT_PGM
Table avec FK :	DEVL	DBAGEST	TRI_PROGRAMME	CD_SECT_PGM	

Analyser

Figure 6.4 Analyse de relation avec SAT : «choix des champs».

L'autre procédure analyse la possibilité de placer une clé étrangère entre deux tables. À la figure 6.4, on indique les critères d'analyse : le S.I.D. d'Oracle, le schéma, les tables, le champ commun sur lequel l'analyse sera exécutée et s'il y a lieu, une description de type texte du champ commun pour ne pas voir que des valeurs numériques ou alphanumériques de la clé primaire.

Lorsque l'on clique sur le bouton [Analyser] (figure 6.4), le résultat de l'analyse s'affiche. La figure 6.5 montre le résultat d'une analyse.

	DBAGEST.TRI_TB_SECTEUR_PGM@DEVL			DBAGEST.TRI_PROGRAMME@DEVL		
Ligne	CD_SECT_PGM	DESC_SECT_PGM	Fréquence	CD_SECT_PGM	Fréquence	
1	01	Brevet d'enseignement	1	01	26	X
2	02	Études préuniversitaires	1	02	1	X
3	03	Libre en recherche - niveau collégial	1	03	1	X
4	04	Périuniversitaire	1	04	1	X
5	1A	Programme court - EIF (9983)	1	1A	4	X
...	...	...	...	...	...	
53	41	Doctorat Honoris Causa	1	41	1	X
54	1D	*Programmes courts (1 session)*	1			G
55	2B	2e cycle - programme extensionné (DESS)	1			G
56	3C	3e cycle Doctorat extensionné UQAM	1			G
57	3D	3e cycle Doctorat conjoint	1			G
58	31	3e cycle - Diplômé	1			G

Analyse de la table avec la clé primaire (PK)			
	Fréq.	**%**	
Nombre de valeurs différentes	58	100%	
Valeurs non utilisées	5	9%	G

Analyse de la table avec la clé étrangère (FK)			
	Fréq.	**%**	
Nombre de valeurs totales	805	100%	
Valeurs avec correspondance sur PK	805	100%	X
Valeurs inexistantes pour FK	0	0%	D

Figure 6.5 Analyse de relation avec SAT : «analyse finale».

Le résultat de cette procédure d'analyse permet d'afficher trois types de relation : celle dont le champ commun est dans les deux tables (identifié par 'X'), celle dont le champ commun est présent dans la première table uniquement (identifié par 'G') et finalement celle dont le champ est présent dans la seconde table (identifié par 'D' mais non présent dans la figure 6.5).

Le tableau sommaire de la fin de la procédure de la figure 6.5 est très important. Il nous indique dans la partie du haut, la fréquence des valeurs de la première table (58 => 100%) et le nombre de valeurs non utilisées de la première table (5 =>9%). La partie du bas nous indique la fréquence de la valeur dans la seconde table (805 => 100%), le nombre de valeurs correspondant entre les deux tables (805 => 100%) et finalement, s'il y a lieu, les valeurs manquantes des deux côtés ('G', et 'D'). Dans cet exemple, une clé étrangère peut être activée puisque toutes les valeurs de la seconde table correspondent avec au moins une valeur de la première table.

L'outil SAT permet d'analyser les tables sur plusieurs S.I.D. Oracle, sur différents schémas de la base de données. De par sa nature, il permet de comparer le contenu de deux tables sur différents S.I.D. afin d'ajuster le contenu des deux tables. Cet outil sera utilisé non seulement pour l'entrepôt de données mais pour les systèmes OLTP existant. L'outil permet aussi d'avoir plus d'un champ commun.

6.1.3 Outils d'analyse des DDL (OAD)

Pour résoudre la dernière problématique de modification de structure, le logiciel OAD, a été créé à l'UQTR. Il permet de lire les journaux (*logs*) de toutes les transactions d'Oracle. Cen fait les fichiers pour les «REDOLOG». De la liste des transactions, sont extraites seulement les commandes DDL (le type d'extraction est paramétrable). L'analyste responsable de la structure de l'entrepôt de données consulte ses commandes et décide s'il doit réagir du côté des structures de l'entrepôt.

Dans un premier temps, la lecture des «logs» journaliers d'Oracle est emmagasinée dans les tables du système OAD comme le montre la figure 6.6. Il existe deux cahrgement. Celui du 2 novembre et celui du 31 octobre. L'administrateur peut consulter et effacer chaque chargement.

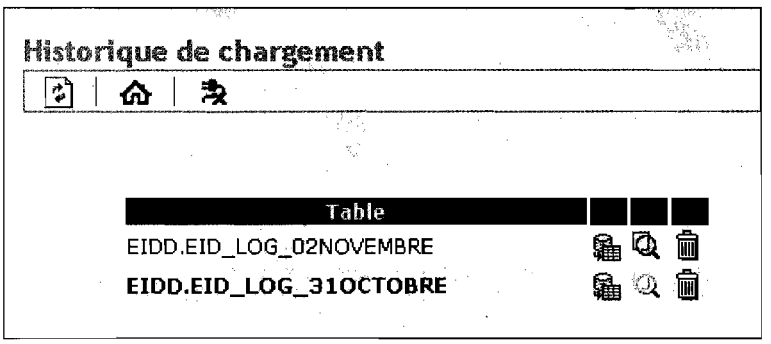


Figure 6.6 Historique des chargements des «logs» d'Oracle.

Dans les *logs* d'Oracle chargés, seules les commandes DDL sont filtrées et emmagasinées pour traitement. En cas d'erreur, on peut recommencer le chargement manuellement. On peut visualiser à la figure 6.7 la liste des commandes DDL du chargement du 1er novembre 2007. Il existe des paramètres de filtrage supplémentaire que l'on peut appliquer aux données qui s'affichent (choix de table, schéma, date et/ou message).

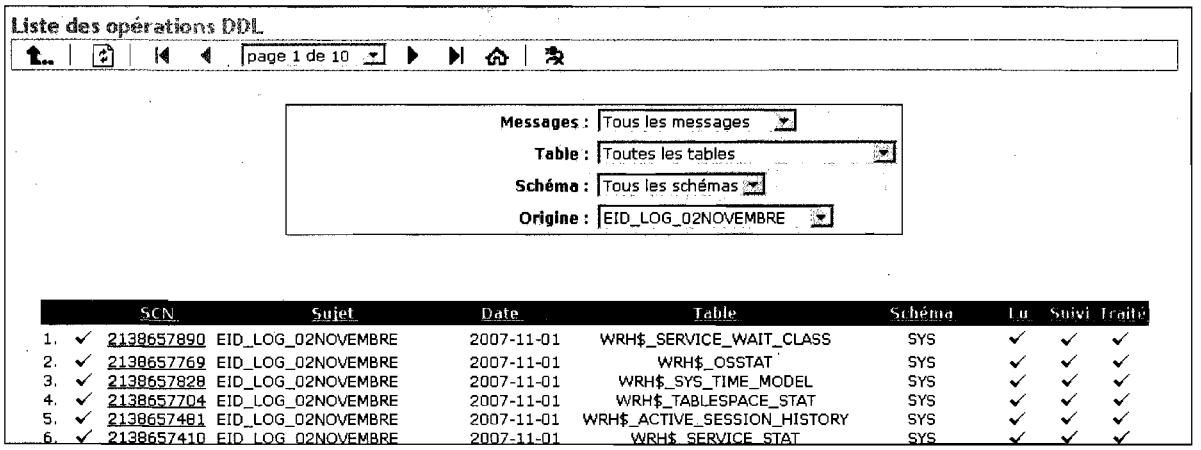


Figure 6.7 Liste des modifications de structure.

Chaque commande DDL est représentée une ligne du résultat de la figure 6.7. La commande, c'est-à-dire son code SQL est visible (figure 6.8) en cliquant sur le numéro de transaction unique situé au début de la ligne (figure 6.7) dont le titre de colonne est «SCN».

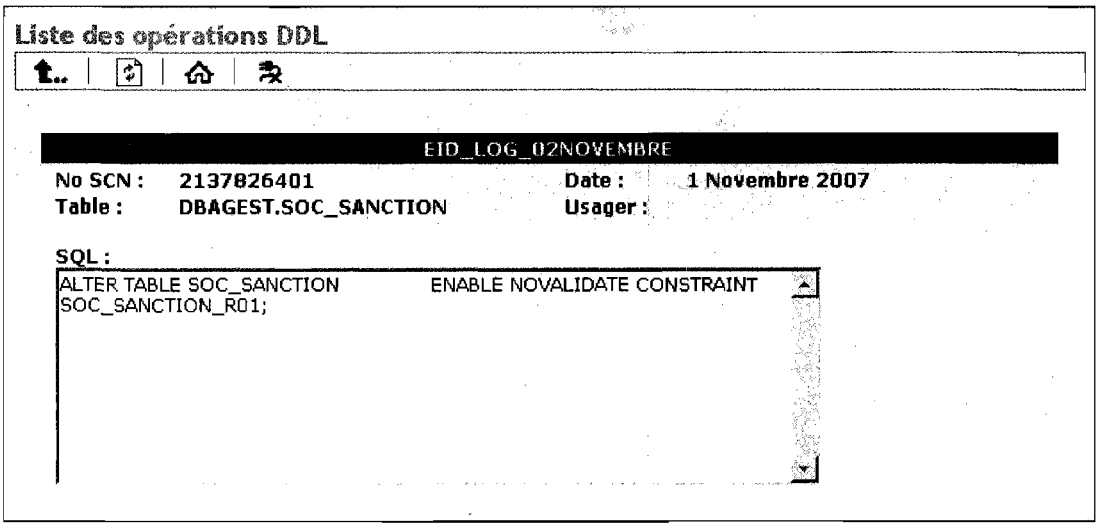


Figure 6.8 La commande DDL en détail.

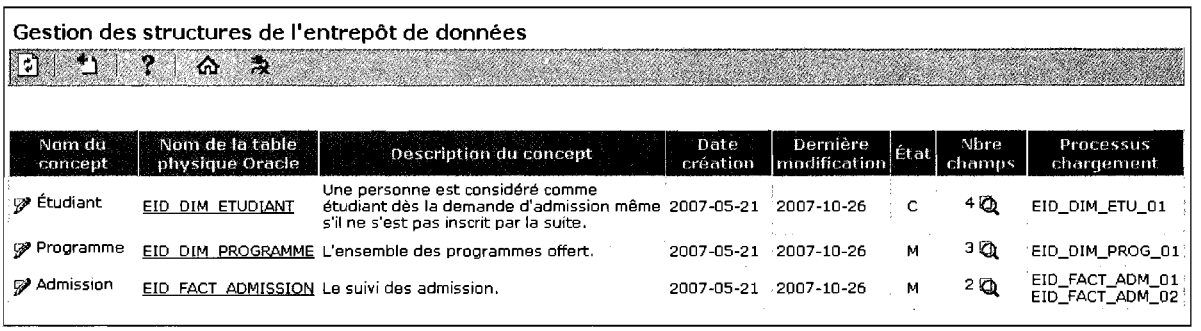
En visualisant la commande, l'administrateur de l'entrepôt peut décider de supprimer cette commande de la liste des DDL en cours car il n'y aura rien modifier du côté de l'entrepôt, de faire une modification dans les structures de l'entrepôt ou de mettre à jour ses métadonnées.

6.1.4 Historisation de la structure

Nous avons créé le logiciel SEPTS (SystèmE de Perception Temporelle des Structures) qui permet de créer l'historisation des structures. Pour le moment, aucun logiciel existant ne permet d'intégrer cette information aux outils BI mais si un jour cette passerelle voyait le jour, nous aurions une longueur d'avance car on pourrait informer les gestionnaires des modifications physiques à travers le temps d'une table et son contenu. Cela permettra d'expliquer aux gestionnaires qu'une donnée est vide pour une période donnée. . Voici les deux grands rôles de ce logiciel :

- Permet l'historisation des structures dimensionnelles de l'entrepôt, et le suivi des processus afin d'informer les gestionnaires des événements qui peuvent influencer les résultats des tableaux de bord.

- Permet la définition des concepts TABLE et CHAMP pour les utilisateurs permettant d'inclure un glossaire en ligne au projet EID.



Nom du concept	Nom de la table physique Oracle	Description du concept	Date création	Dernière modification	État	Nbre champs	Processus chargement
Étudiant	EID_DIM_ETUDIANT	Une personne est considéré comme étudiant dès la demande d'admission même s'il ne s'est pas inscrit par la suite.	2007-05-21	2007-10-26	C	4	EID_DIM_ETU_01
Programme	EID_DIM_PROGRAMME	L'ensemble des programmes offert.	2007-05-21	2007-10-26	M	3	EID_DIM_PROG_01
Admission	EID_FACT_ADMISSION	Le suivi des admission.	2007-05-21	2007-10-26	M	2	EID_FACT_ADM_01 EID_FACT_ADM_02

Figure 6.9 Historisation des structures.

La figure 6.9 nous montre le nom des concepts existant, leur définition, le nom physique des tables, les dates de création et de modification, l'état de la table, la liste des champs de la table et ses processus de chargement.

6.2 Modélisation dimensionnelle

La création du modèle dimensionnel se fait en 4 étapes, selon le cycle de vie décisionnel de Kimball que voici :

- Identifier le processus d'affaires,
- Identifier la granularité de la table de faits,
- Identifier les dimensions,
- Identifier les faits.

Le processus d'affaires choisi est «le suivi de l'admission». La modélisation du processus d'affaires est effectuée à l'aide de l'arbre du sujet à la figure 6.10.

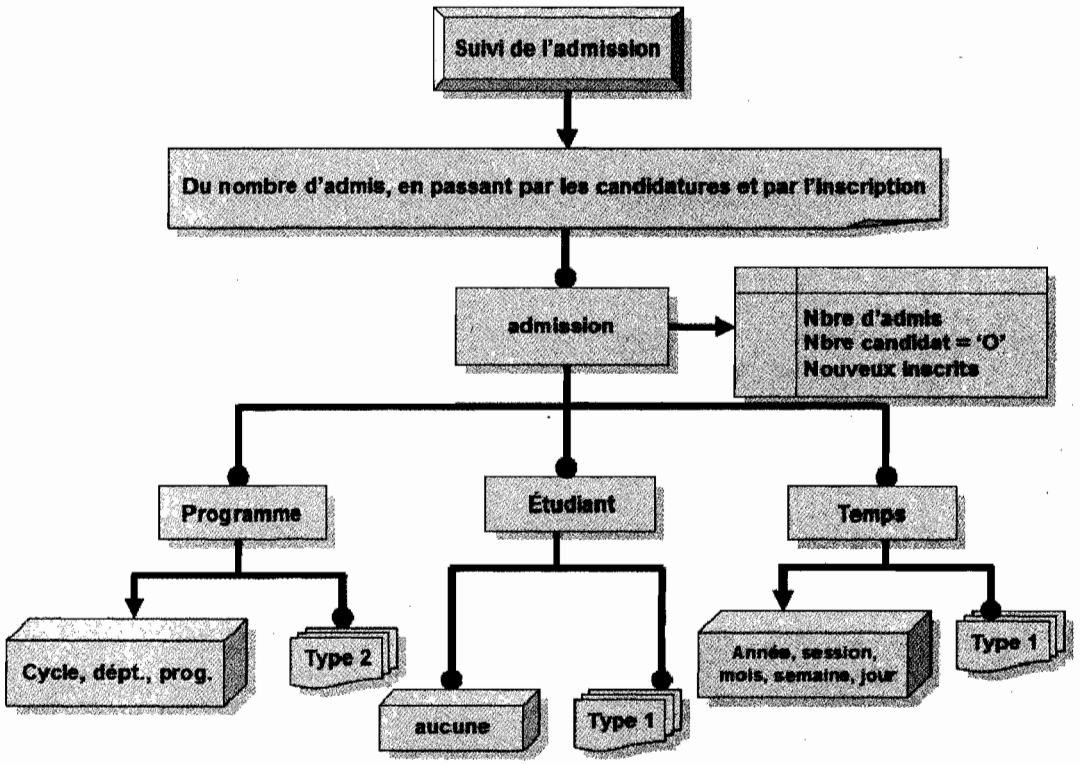


Figure 6.10 Arbre du sujet de l'admission.

6.3 Modèle physique de l'entrepôt

En appliquant les techniques de dénormalisation, le schéma entité-relation de la figure 6.2 sera converti en schéma dimensionnel à la figure 6.11 et sera par la suite, créé physiquement.

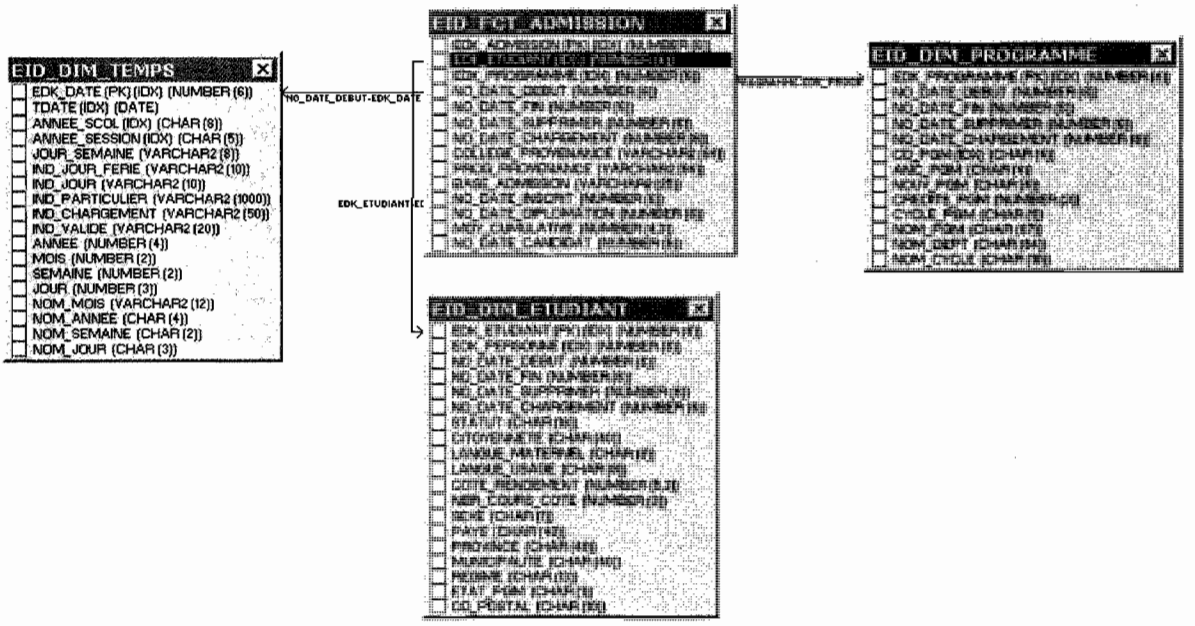


Figure 6.11 Le modèle dimensionnel de l'admission.

Si l'on compare la figure 6.2 à la figure 6.11, on peut remarquer qu'il y a beaucoup moins de tables, et ce, en sachant que toutes les données requises font partie des deux schémas. Puisqu'il y a moins de tables, il y aura moins de jointures et un gain considérable de performance. Un autre avantage non négligeable est la compréhension intuitive du modèle dimensionnel pour l'utilisateur.

6.4 Extraction, transformation et chargement de l'entrepôt

Présentement la phase ETC reste à définir. Les huit phases du processus ETC de Monsieur Rémy Choquet pourront être appliquées prochainement lors de l'installation de la suite de Business Object (BO).

Cependant, deux méthodes de pompage ont été testées soient les vues matérialisées et le CDC d'Oracle.

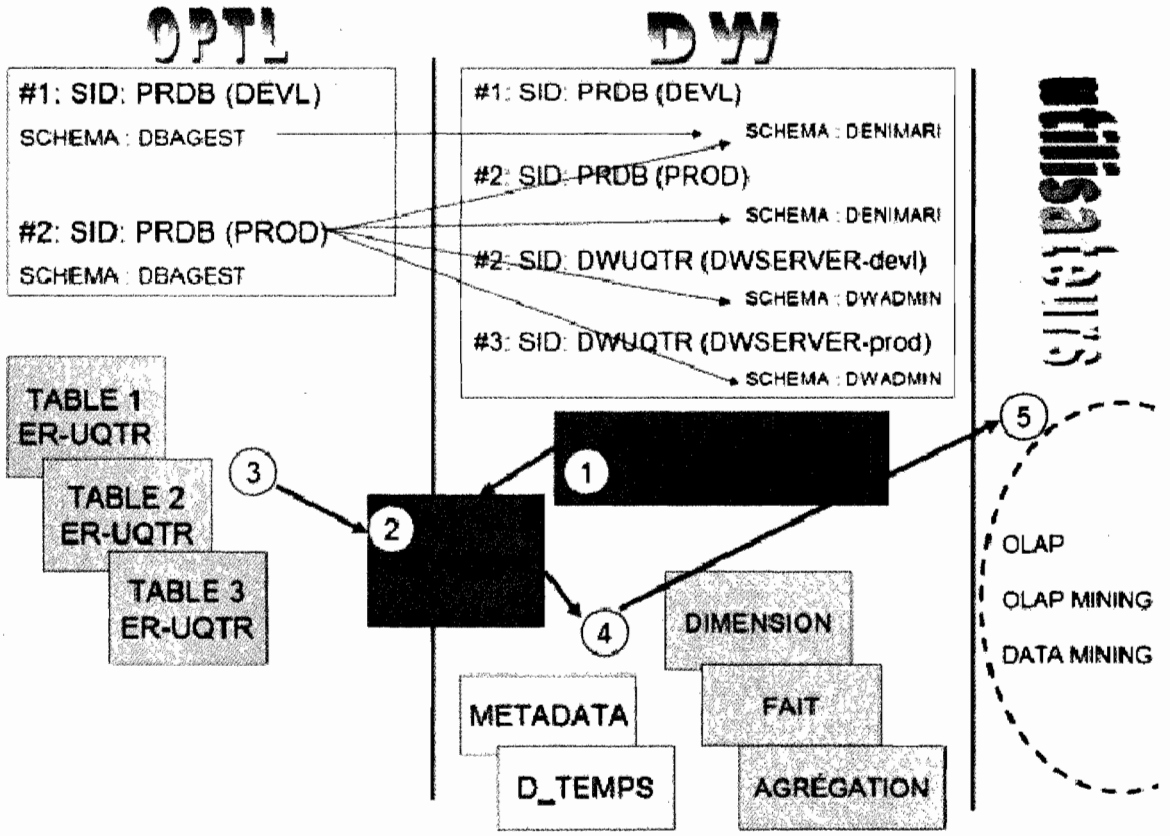


Figure 6.12 Chargement avec les vues matérialisées.

La réalisation de la méthode de chargement avec les vues matérialisées, schématisée à la figure 6.12, est déroulée en 5 phases. Comme pré requis, il faut choisir le schéma Oracle (#1 :SID : PRDB (DEVL)) sur lequel on va travailler. Les données de ce schéma seront les sources de l'entrepôt. Il faut aussi choisir le schéma Oracle vers lequel les données seront emmagasinées. Ce schéma conservera les données cibles du chargement. Il faut créer les structures physiques (p.ex. : tables, champs, clés) dans le schéma cible du côté de

l'entrepôt avant le chargement. Par la suite, les vues matérialisées associées sont créées. Pour le moment, les tables physiques des vues sont vident. Une fois les données à charger déterminées et les vues créées, l'on peut débuter le processus de chargement.

À la phase #1 de la figure 6.12, la création de modules de programmation en PL/SQL permet de configurer manuellement ou automatiquement selon une fréquence donnée les vues matérialisées. À la phase #2, les *logs* des vues matérialisées sont activés et attendent les changements qui seront effectués sur données sources. À la phase #3, aussitôt qu'un ajout, qu'une modification ou qu'une suppression est effective sur les données sources, les *logs* des vues matérialisées associés accumulent ces modifications, et ce, séquentiellement.

La phase #4 permet de charger l'entrepôt. Si la vue est automatiquement rafraîchie tous les jours, à minuit, les données seront transférées du côté de l'entrepôt par des processus programmés en PL/SQL. Ces programmes alimenteront les tables de faits, les tables de dimensions, les agrégats et les métadonnées. Toutes ces données seront estampillées d'une donnée supplémentaire qui représentera la dimension temporelle.

La dernière phase de la figure 6,12 permet d'extraire les données de l'entrepôt sous différentes formes pour les analyser. Il sera possible de faire de l'OLAP et/ou du *data mining*.

Cette méthode ne nous convenait pas puisqu'il y avait une possibilité de perte d'information entre le moment où la vue est rafraîchie et que l'on charge l'entrepôt. Entre ces deux opérations, l'entrée de nouvelles modifications pouvait être perdue. De plus, la modification d'une donnée est encodée dans un vecteur et non accessible directement en clair pour un traitement. Par exemple, si je modifie l'adresse d'une personne ayant le numéro de membre #254323 alors le *log* de la vue captera qu'il y a eu une modification de la clé primaire #254323, mais on ne saura pas directement qu'il s'agit d'un changement d'adresse.

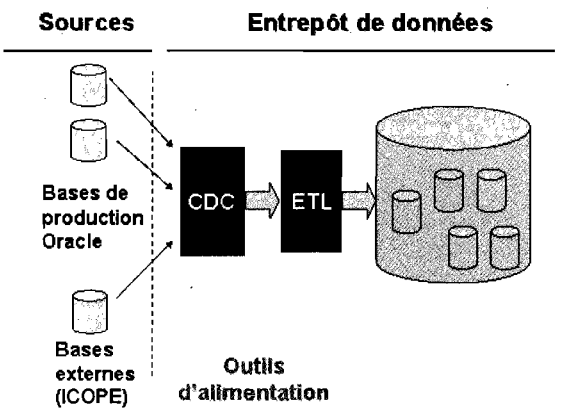


Figure 6.13 Chargement avec le CDC d'Oracle.

Comme le montre la figure 6.13, l'emplacement de la zone temporaire de traitement des données CDC se positionne entre les données sources et l'outil de chargement ETL. Le module d'Oracle a été installé pour acquérir toutes modifications aux données sources qui étaient prédéterminées au chargement de l'entrepôt.

Dans un premier temps, les tables servant de sources pour lesquelles on veut surveiller les changements sont sélectionnées. Le module CDC d'Oracle capture les commandes DML et les insère dans une table de modifications (change table). Nous avons sélectionné les tables du système du «dossier étudiant» pour la partie de l'admission dont le nom débute par «ADM_» ainsi que celles débutant par «DAF_». Les *changes tables* s'alimentent des modifications des tables sources avec le *package* DBMS_LOGMNR_CDC_PUBLISH. Selon une certaine fréquence, on effectue un «select» pour récupéré les modification des *changes tables* à l'aide du package DBMS_LOGMNR_CDC_SUBSCRIBE et on applique l'ajout vers l'entrepôt. Lorsque l'opération est réussie, il suffit de vider les *changes tables* et de les remplir à nouveau.

Cette méthode est très rapide. Elle n'affecte presque pas le temps de réponse des systèmes transactionnels et permet de lire en texte clair le changement effectué sur une donnée. C'est cette méthode qui sera utilisée en amont de l'ETC de BO.

Présentement, le chargement est fait à partir de programmes PL/SQL. L'outil ETC de BO est attendu avec impatience.

Chapitre 7
Publication des données

Dans ce chapitre, les étapes de la publication des données seront abordées. Une fois les données publiées, l'entrepôt de données sera prêt à répondre aux requêtes des usagers et il sera enfin possible de construire un premier tableau de bord. L'ampleur du travail réalisé précédemment simplifie la phase de présentation des données. On peut comparer les tableaux de bord obtenus pour les dirigeants à la pointe visible de l'iceberg. Simples à première vue, mais toute la simplicité réside dans la face cachée du travail minutieux accompli. Ce chapitre est divisé en trois points :

7.1 Publication des données

On revient sur les cinq rôles de l'entrepôt (voir tableau 3.2). Les trois premiers rôles : prise de données, intégration et distribution, permettent d'alimenter l'entrepôt de données et de s'assurer de la validité et de la qualité des données. Après un chargement qualifié comme «réussi», les données sont alors publicisées, c.-à-d. prêtes à être exploitées par les logiciels de «*reporting*», disponibles aux usagers.

Ce service de livraison rend accessible les bonnes données aux bons utilisateurs. C'est après l'application des règles de sécurité et d'accès que l'utilisateur ayant les droits, pourra consulter et extraire les données ainsi disponibles.

La phase de publication des données ainsi que les règles de sécurité d'accès d'application ne sont pas abordées en profondeur par les fabricants de produits. Par contre, un point commun semble s'en dégager notamment l'application du protocole LDAP pour la sécurité d'accès.

En résumé, le protocole LDAP définit comment s'établit la communication client-serveur. Il fournit à l'utilisateur des commandes pour se connecter ou se déconnecter, pour rechercher, comparer, créer, modifier ou effacer des entrées. Des mécanismes de chiffrement (SSL ou TLS) et d'authentification (SASL), couplés à des mécanismes de règles d'accès (ACL) permettent de protéger les transactions et l'accès aux données.

La plupart des logiciels serveurs LDAP proposent également un protocole de communication serveur-serveur permettant à plusieurs serveurs d'échanger leur contenu et de le synchroniser (replication service) ou de créer entre eux des liens permettant ainsi de relier des annuaires les uns aux autres (referral service).

Source : http://www-sop.inria.fr/semir/personnel/Laurent.Mirtain/ldap-livre.html#I

7.1.1 Indicateurs correctifs d'analyse des données

«S'assurer de la validité et de la qualité des données»

Une fois les données livrées, publicisées, les demandeurs interrogent via des requêtes et des interfaces logicielles les données disponibles dans l'entrepôt. Il faut s'assurer qu'une même requête donne toujours les mêmes résultats. La confiance des utilisateurs envers n'importe quel système peut être perdue si une erreur de taille survient, ou si différents rapports donnent différents résultats alors qu'on aurait dû s'attendre au même résultat. En s'assurant de l'exactitude des données, on s'assure de l'exactitude des résultats. La confiance envers l'entrepôt de donnée est très fragile, mais si elle est acquise, elle sera alors totale et absolue. Si cette confiance est affaiblie, le projet entier peut être remis en question. Un demandeur peut interpréter les données à sa façon, de ce fait, on n'y peut rien, mais les données doivent toujours être les «bonnes données».

Des problématiques partiellement résolues au chapitre 6, nous ne devons pas que livrer la marchandise aux clients, il faut l'informer des particularités du produit pour ne pas qu'il retourne la marchandise. Pour s'assurer de la validité et de la qualité des données, le demandeur se doit d'être informé des faits et des événements pouvant affecter l'interprétation d'un résultat.

Deux volets doivent être couverts. Le premier correspond aux événements qui se sont produits et qui peuvent modifier l'interprétation des données, par exemple, la grève des chargés de cours. Le deuxième est de fournir l'information temporelle des modifications de structure qui expliquerait par exemple, dans une répartition du nombre de diplômés dans un cheminement normal, un pourcentage élevé d'étudiants non diplômés sur les 35 dernières années si la date de diplômation existe seulement depuis 1980. Une proposition d'information aux demandeurs est faite à la section 7.3.

7.2 Présentation des données

L'entrepôt de données permet aux décideurs d'avoir accès plus rapidement et plus facilement aux données stratégiques de l'Université. Les données étant présentes dans les systèmes OLTP, les décideurs pouvaient extraire tout de même des données à des fins d'analyse, mais le processus était compliqué, fastidieux et pas vraiment accessible directement par le demandeur. Il fallait passer par le service de l'informatique soit le SSPT.

Dans les dernières années, voici comment on obtenait des résultats pour l'analyse : une analyse des besoins sur les données était effectuée, suivie de l'élaboration d'une requête SQL et finalement, les résultats de cette requête étaient redirigés aux demandeurs sous le format d'un fichier Excel.

Par la suite, un logiciel paramétrable d'extraction en ligne a permis d'automatiser les mêmes demandes d'extraction de façon dynamique. C'est le logiciel «Discoverer Viewer» d'Oracle qui était utilisé et est encore utilisé présentement. Les requêtes se devaient d'être conçues tout de même par le SSPT.

Certains utilisateurs, plus près des savoirs de l'informatique, avaient accès directement aux structures des données et concevaient eux-mêmes leurs requêtes. Ces utilisateurs sont peu nombreux. Le logiciel «Discoverer Desktop» permet à un utilisateur «expert» de construire ses demandes d'extraction. Les résultats sont disponibles par la version «Desktop» ou la version «Viewer» de «Discoverer» dépendamment si l'utilisateur veut partager ses requêtes avec d'autres personnes.

Les deux versions de «Discoverer», le «Desktop» ou le «Viewer», reposent sur l'inclusion par le logiciel «Discoverer administrator» de structures pouvant être accessibles par les usagers. L'administrateur de ce dernier logiciel est un analyste du SSPT.

Actuellement, en fonction des demandes de plus en plus nombreuses des usagers, du volume élevé de données dans les résultats demandés et de la complexité des demandes, le logiciel «Discoverer» a atteint ses limites. Certaines demandes d'extraction ont comme réponse «*Application Server not response*», indiquant que le processus d'extraction est trop long et qu'aucune donnée n'est présentée au demandeur.

On peut sentir l'évolution des systèmes décisionnels. Auparavant, comme le montre le graphique de gauche de la figure 7.1, seuls quelques utilisateurs avaient besoin de données et maintenant, en consultant le graphique de droite, chaque niveau hiérarchique de l'entreprise a un besoin d'extraction, différent certes, mais nécessaire à l'accomplissement de leurs tâches et leurs prises de décision.

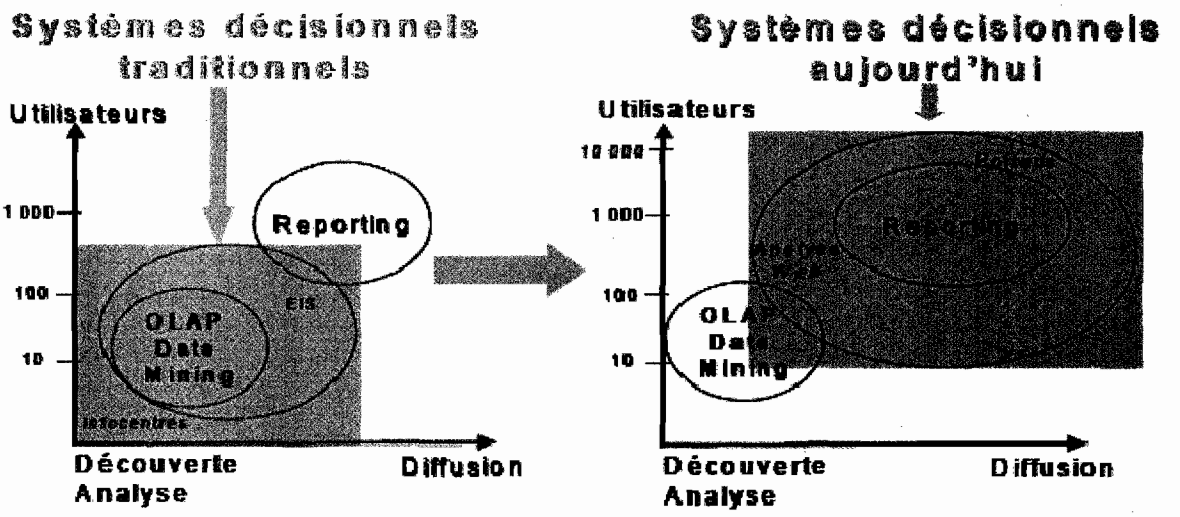


Figure 7.1 ***Évolution des systèmes décisionnels.***
(Source : Anne-Marie Abisségué, IDC. 2004, France)

7.2.1 Préparation du moteur «ROLAP» pour l'extraction des données

Le service «présentation des données» est la seule porte de sortie des données de l'entrepôt. Afin de rendre accessibles les données aux utilisateurs, il faut, tout comme «Discoverer Administrator», choisir au préalable les structures qui seront accessibles par les usagers. Chaque outil de présentation des données communément appelé outil de «reporting» possède sa propre interface d'inclusion des structures afin de les rendre accessibles aux utilisateurs par le «reporting».

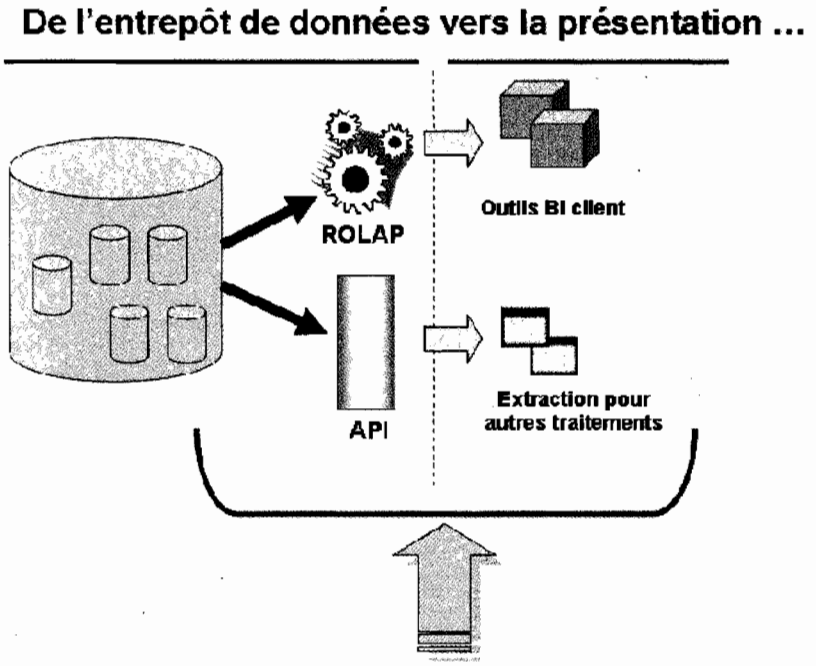


Figure 7.2 Service de présentation des données.

L'outil de «reporting» qui sera choisi aura un double mandat. Permettre l'extraction de données de l'entrepôt mais aussi puisque aucun outil de «reporting» convivial n'est disponible pour les usagers, être utilisé directement sur les systèmes transactionnels.

La compagnie COGNOS est venue installer sa plate-forme de «reporting» au SSPT afin de permettre au Service de tester leurs outils dans une phase d'exploration et d'évaluation et ce, pour l'extraction de l'entrepôt de données.

Il est essentiel avant de construire un tableau de bord de connaître les mesures qui seront représentées. Ces mesures représentent les indicateurs de performance. Après l'analyse de ces mesures par le gestionnaire, il déterminera s'il faut agir, réagir ou conclure.

Il est maintenant temps de réaliser l'inclusion des structures au moteur «ROLAP» de COGNOS pour les tableaux de bord. L'outil COGNOS correspondant se nomme le «FRAMEWORK MANAGER». On peut voir à la figure 7.3, la fenêtre de la vue du projet (project viewer) de ce logiciel.

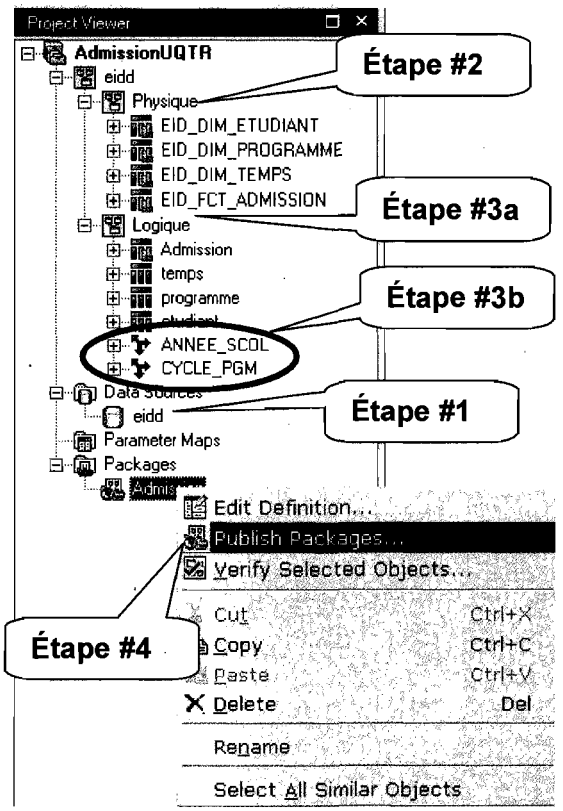


Figure 7.3 La vue d'un projet du «Framework manager de COGNOS».

Étape #1 : *Créer le branchement à la source de données.*

Cette étape se réalise très rapidement. Il suffit d'avoir créé un nom de liaison pour le serveur Web au SID de la base de données de l'entrepôt.

Étape #2 : *Choisir les structures que l'on veut inclure (espace physique).*

Une fois la source de données définie, et la connexion testée et réussie, un espace de travail nommé «physique» est créé. Ensuite, une liste des objets existants de la base de données apparaît. Il suffit de cocher les tables, vues, fonctions ou autres objets que l'on désire inclure au modèle. Les relations entre les objets seront définies dans cet espace. Seul l'administrateur du «Framework manager» pourra modifier cet espace de travail.

Il est important à cette étape de définir le type des attributs de chaque table afin d'en indiquer les mesures, les attributs et les clés.

Étape #3 : *a) Choisir les structures que l'on veut publier (espace logique).*

Pour plus de contrôle et de sécurité, un espace logique est créé. C'est cet espace qui sera visible par les usagers. Cet espace ne contiendra pas les relations physiques du modèle. L'usager ne pourra donc pas les modifier ni les supprimer. Ce contrôle évitera bien des soucis de performance, par exemple, si une relation était supprimée. S'il n'existe pas de relation entre deux tables, un produit cartésien se réalisera lors de la jointure, un désastre pour la performance, et l'utilisateur constatera, mais peut-être pas, la présence de doublons, triplons, centrions ..., qui rendra nuls les résultats des analyses. L'utilisateur, dans la vue logique, pourra par contre se créer de nouvelles relations.

Dans cet espace, on peut exclure certaines tables ou certains champs du modèle pour ne pas les offrir aux utilisateurs. On applique ici les premières règles de sécurité. On pourra créer plusieurs espaces logiques et permettre à différents utilisateurs de voir différentes données. La figure 7.4 montre l'espace physique et l'espace logique.

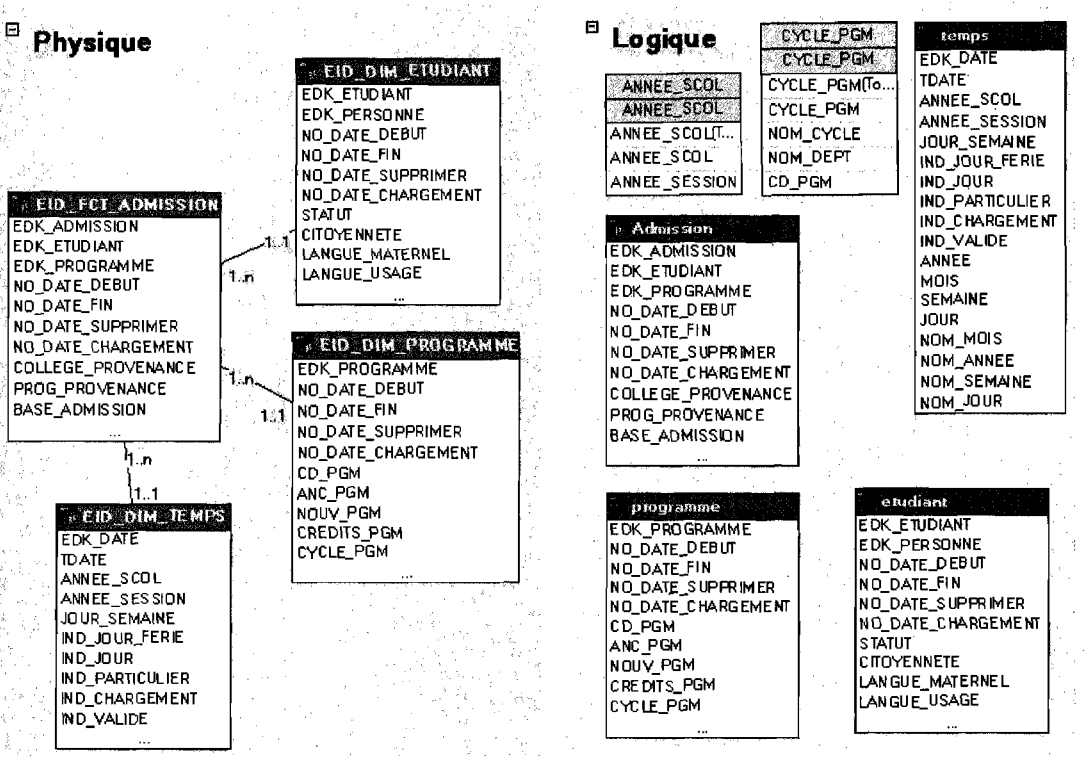


Figure 7.4 L'espace physique et l'espace logique du «Framework manager».

b) Déterminer la hiérarchie des dimensions.

Dans la vue logique, on déterminera les niveaux hiérarchiques des données d'une dimension. Cette fonctionnalité permettra aux utilisateurs d'accéder rapidement à une vue plus ou moins en détail de l'ensemble des données. Par exemple, on pourra voir le nombre d'admis par cycle, le nombre d'admis par département ou le nombre d'admis par programme. On comprend, à la figure 7.5, que les programmes sont inclus dans les départements qui eux appartiennent à un cycle.

Dimensions	
ANNEE_SCOL	CYCLE_PGM
ANNEE_SCOL	CYCLE_PGM
ANNEE_SCOL(Tout)	CYCLE_PGM(Tout)
ANNEE_SCOL	CYCLE_PGM
ANNEE_SESSION	NOM_CYCLE
	NOM_DEPT
	CD_PGM

Figure 7.5 Hiérarchies des dimensions.

Étape #4 : *Publication des données.*

Une fois les choix finaux effectués, on cré un ensemble de publication «package» qui sera visible par les utilisateurs. Cet ensemble rend accessible seulement l'espace logique (c'est par choix uniquement, que l'on peut aussi rendre accessible l'espace «physique»). Il suffit de cliquer sur «Package», et d'utiliser l'option «Publication du package» sur le nom de l'ensemble désiré. Les données sont présentement disponibles aux usagers ayant les droits de les voir. La durée de la réalisation de ces quatre étapes a été de trente minutes.

7.2.2 Tableaux de bord

Une fois les données publiées, l'utilisateur, par des outils logiciels de «reporting», construira ses tableaux de bord. Pour la suite COGNOS 8, les outils sont «Query Studio», «Analyse Studio» et «Report Studio». L'outil «Query Studio» permet des requêtes simples mais rapides pour les tableaux de bord. L'outil «Analyse Studio» permet la création et l'exploration des cubes de données. L'outil «Report Studio» permet la création de tableaux de bord plus sophistiqués.

La différence entre un cube de données d'«Analyse Studio» et un tableau de bord de «Report Studio» repose sur la possibilité de changer la vue de l'utilisateur sur les données comme il le désire et la rapidité de navigation entre les dimensions et leurs granularités. Le cube de données permet l'intégration de pivots laissant libre choix à l'usager de considérer une vue des données par rapport à une autre. Tandis que dans le tableau de bord, il faut créer chaque rapport manuellement et faire des liens entre les différentes vues. Dans la nouvelle version de «Report Studio», un outil «pivot» est présent. Une exploration de cette nouvelle fonctionnalité permettra peut-être de combiner les deux.

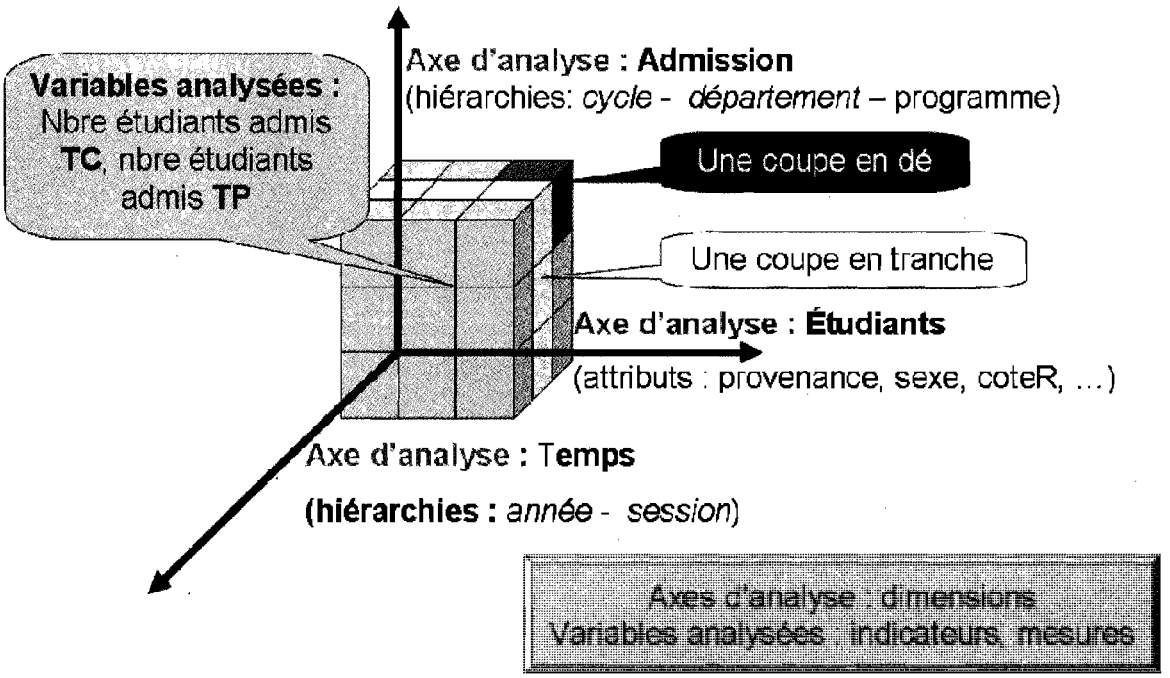


Figure 7.6 Le cube de données.

La figure 7.6 représente les données dans un cube. Trois axes représentent les dimensions et les faits (les mesures) sont les variables que l'on trouve sur les différentes coordonnées spatiales du cube (vision 2D ou vision 3D).

Lorsque l'on coupe un cube sur un axe où l'autre on fait une opération sur les tranches du cube. Cette opération se nomme «slicing». C'est une coupe en tranche (voir figure 7.6). Cela représente tous les étudiants de la session d'hiver 2007 (2007-01).

Si l'on navigue à travers les dimensions et que l'on coupe une tranche pour obtenir un sous-groupe de données on appelle cette opération «dicing». C'est une coupe en dé (voir figure 7.6). Les données pourrait être tous les étudiants de la session d'hiver 2007 (2007-01) admis au 1er cycle.

Les outils OLAP permettent de traiter rapidement les changements de vue de l'utilisateur sans avoir à recréer le cube. Il y a une multitude de rapports dans un seul cube de données.

Il est maintenant le temps de réaliser un premier tableau de bord. L'outil «Report Studio» sera utilisé. Dans la modélisation de l'arbre du sujet, deux mesures ont été définies. La première étant le nombre d'admissions par session par programme et la deuxième, le nombre de candidats qui se sont inscrits par la suite à au moins une session. La demande de monsieur Rémy Auclair permet de construire le premier tableau de bord permettant d'offrir une vue d'ensemble sur le suivi de l'étape d'admission. Il indiquera quels sont les étudiants qui se sont inscrits, à quelle session et leur régime d'études.

Étape #1 : *Construire le modèle du tableau de bord en mode gestionnaire.*

L'interface de «Report Studio» tel que vu à la figure 7,7 est assez simple, elle se divise en trois zones. La zone des données, la zone des propriétés et la zone du rapport. La zone du rapport peut être personnalisée à souhait. Pour le tableau de bord, un en-tête et un pied de page ont été définis. Par la suite, un «glisser/déplacer» permet d'ajouter les champs désirés de la zone des données à la zone du rapport.

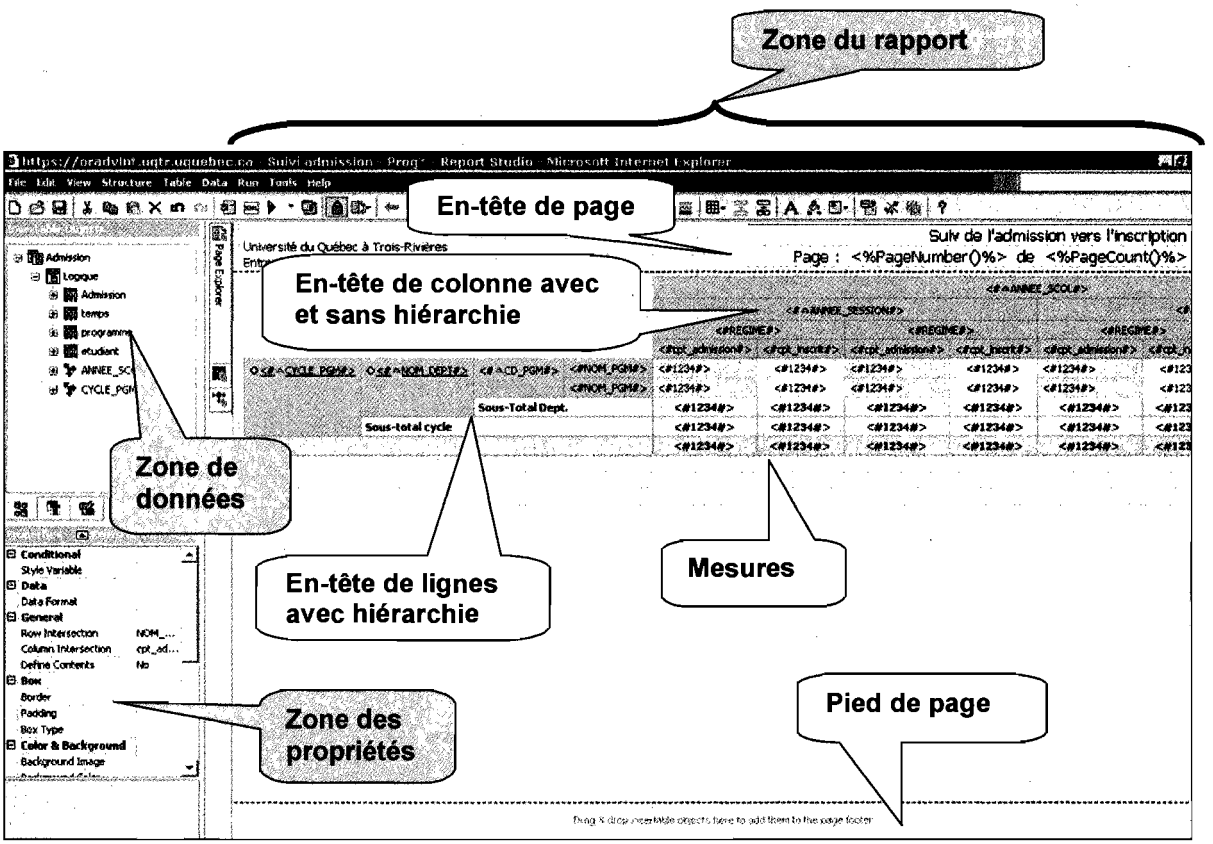


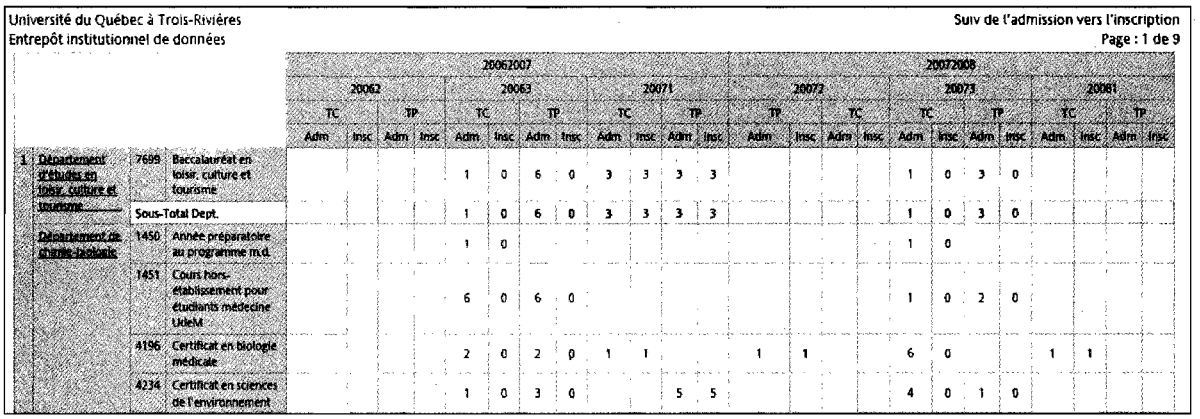
Figure 7.7 L'interface du «Report Studio».

Une fois, les données sélectionnées, il reste l'application de style et de format à effectuer selon les goûts du gestionnaire. Mettre en place les trois rapports (par programme, par département, par cycle) a pris environ une heure trente minutes pour la sélection de données et trente minutes pour la mise en page. Pour un total de deux heures en incluant la mise en place de l'espace «logique».

Pour obtenir le même rapport avec «Discoverer Web», incluant aussi la définition des structures offertes aux usagers avec «Discoverer Administrator», deux jours de travail furent requis. Le gain en performance était indéniable en comparant les schémas relationnel et dimensionnel, maintenant, le gain en ressources humaines l'est tout autant.

Étape #2 : *Exécuter le rapport en mode client.*

On peut par «Report Studio» exécuter la requête. La figure 7.8 nous présente le 1er tableau de bord.



Université du Québec à Trois-Rivières
Entrepôt institutionnel de données

Suiv de l'admission vers l'inscription
Page : 1 de 9

			20062007										20072008							
			20062				20063				20071		20072				20073			
			TC		TP		TC		TP		TC		TP		TC		TP		TC	
			Adm	Insc	Adm	Insc	Adm	Insc	Adm	Insc	Adm	Insc	Adm	Insc	Adm	Insc	Adm	Insc	Adm	Insc
1	Département d'études en loisir, culture et tourisme	7699	Baccalauréat en loisir, culture et tourisme				1	0	6	0	3	3					1	0	3	0
		Sous-Total Dept.					1	0	6	0	3	3					1	0	3	0
	Département de chimie-biologie	1450	Année préparatoire au programme m.d.				1	0									1	0		
		1451	Cours hors-établissement pour étudiants médecine UdeM				6	0	6	0							1	0	2	0
		4196	Certificat en biologie médicale				2	0	2	0	1	1			1	1		6	0	
		4234	Certificat en sciences de l'environnement				1	0	3	0			5	5				4	0	1

Figure 7.8 L'exécution du rapport par «Report Studio».

Il existe un portail COGNOS 8 par lequel les utilisateurs se branchent pour avoir accès aux tableaux de bord. On peut regrouper les tableaux de bord dans différents dossiers par sujet, thème ou service. Ces dossiers représentent les «*packages*» que l'on publie à l'étape 4 de la section 7.2.1. Par des droits de lecture sur ces dossiers, l'utilisateur peut voir les rapports auxquels il a droit. Il exécute la requête du rapport afin d'en consulter les résultats. À la figure 7.9, l'utilisateur choisit le dossier «Admission».

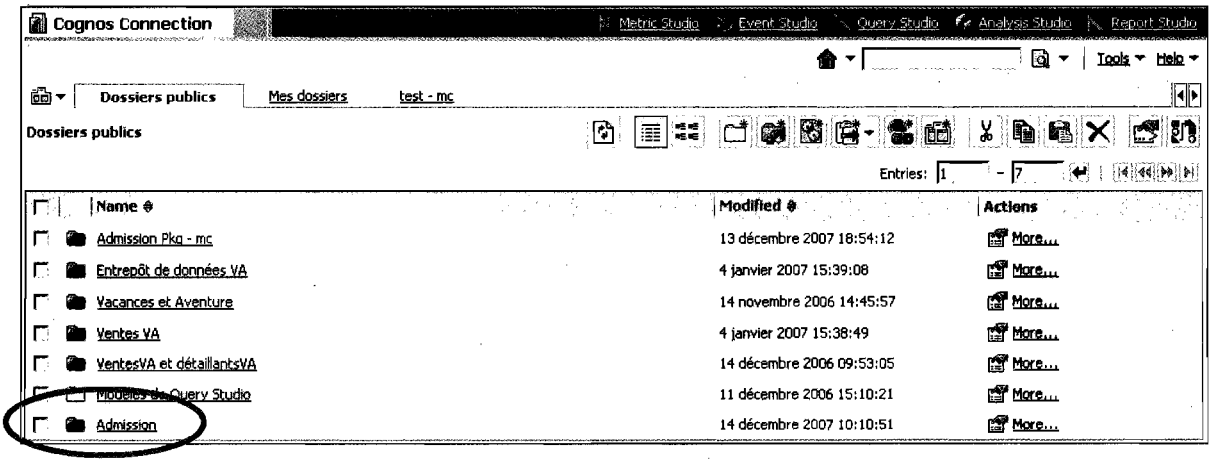


Figure 7.9 L'entrée du portail : choix du dossier.

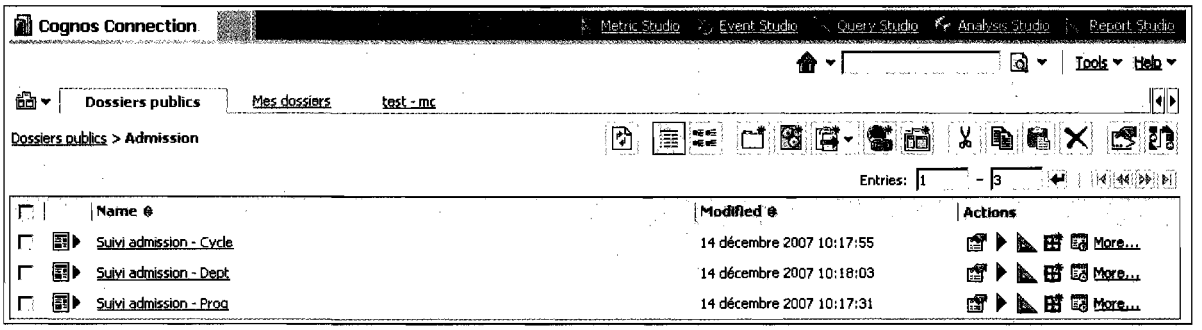


Figure 7.10 Les rapports du dossier choisi.

En cliquant sur le dossier «Admission», la liste des tableaux de bord du dossier s'affiche (figure 7.10). On retrouve les trois tableaux de bord construits à l'étape précédente. L'utilisateur peut consulter les résultats en cliquant sur le nom du rapport.

7.2.3 Forage des données à l'intérieur des hiérarchies des dimensions

Le terme «forage» est utilisé dans le sens de navigation à travers les différentes hiérarchies des dimensions appelées aussi *«drill down».* On peut donc affiner ou agréger les opérations placées sur le rapport. Il ne faut pas interpréter «forage de données» dans le sens du «data mining» qui représente plutôt une technique de fouille permettant d'extraire de l'information complémentaire et des modèles de connaissance explicatifs ou prédictifs.

Il existe deux façons d'explorer la hiérarchie des dimensions. Une avec un pivot par le cube de données de l'outil «Analyse Studio», l'autre en plaçant des liens manuellement entre les rapports. On peut monter ou descendre dans la granularité de la dimension en cliquant sur le lien. À la figure 7.11, un lien a été créé manuellement entre les tableaux de bord. L'utilisateur peut voir les résultats plus ou moins agrégés en modifiant sa vue à travers la dimension. Il pourra voir les résultats en fonction des programmes, par département ou par cycle et naviguer à souhait entre ces trois vues.

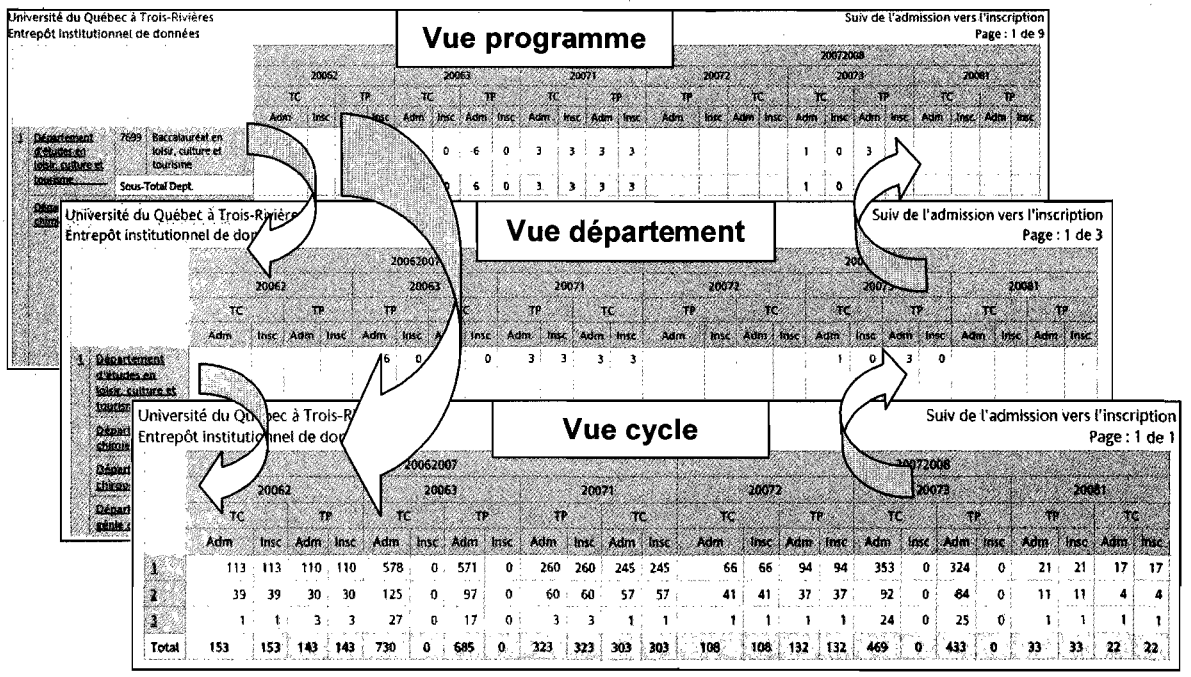


Figure 7.11 Le forage des dimensions.

7.3 Proposition d'une interface d'extraction (API) pour fichiers plats

Les outils de «Reporting» offrent différents formats de sortie : HTML, PDF, Excel, XML et fichier texte avec séparateur (CVS). Nul besoin de créer un outil supplémentaire pour l'extraction de données.

Cependant, voici une proposition d'extraction de «fichier plat» pour le data mining. Les fichiers plats sont des fichiers textuels possédant une structure ordonnée, définie par son créateur. Les données contenues dans ce genre de fichiers peuvent être facilement extraites par des outils d'analyse complémentaire. Chaque ligne d'un fichier est composée d'un ensemble de champs séparé par un délimiteur. Chaque ligne peut être équivalente à un enregistrement d'une table dans une base de données. Le mot «fichier plat» prend tout son sens en le comparant au résultat d'une requête SQL structurée en plusieurs tables. Le résultat de la requête donnera l'équivalent d'une seule table par un ensemble d'enregistrements que l'on peut sauvegarder dans un seul fichier. Ce fichier dit «plat» sera structuré avec les mêmes champs inclus dans la requête initiale. Ce fichier peut contenir énormément de lignes et plusieurs champs par ligne.

Cet outil d'extraction permet à un utilisateur expert de naviguer à travers les concepts de l'entrepôt. Chaque concept est défini à travers l'ontologie des données et possède sa définition dans le domaine. L'utilisateur peut choisir les concepts sur lesquels l'analyse portera. À chaque choix, la définition de l'objet sélectionné s'affiche à l'usager. Le mémoire de maîtrise de ma collègue Yanfen Shen du groupe de recherche en *data mining* [SHEN 07] fourni les explications essentielles à l'ontologie des données. Par une interface API, on peut lire l'ontologie et afficher à l'utilisateur les concepts pour le prototype d'extraction. Le prototype se présente en une seule page séparée dans ce document par les figures 7.12 et 7.13. Le prototype présente l'ensemble des fonctionnalités afin de permettre à un utilisateur d'extraire les données de l'entrepôt en respectant des critères prédéfinis de son choix pour en faire une analyse avec d'autres de type de logiciel comme «WEKA» pour le *data mining*.

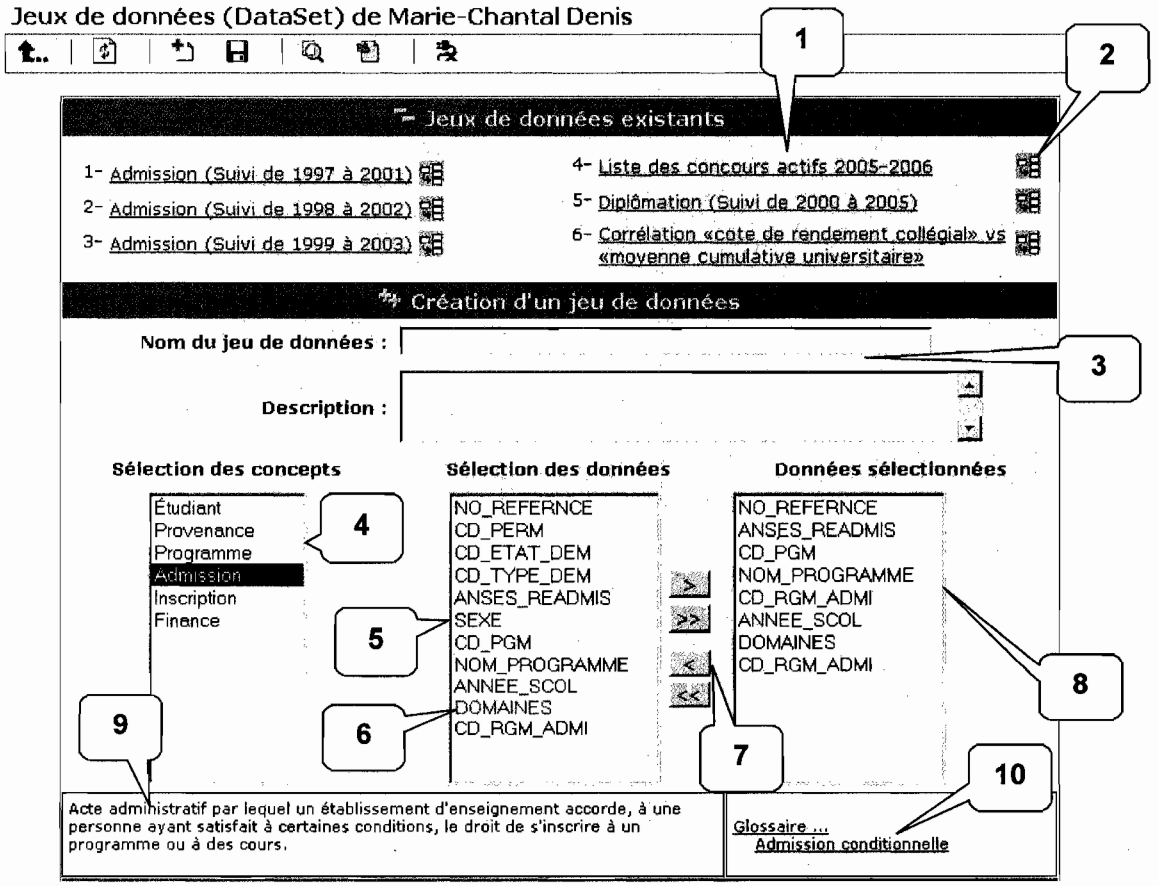


Figure 7.12 Interface d'extraction (API) – partie du haut.

L'utilisateur pourra enregistrer sa configuration d'extraction afin de la modifier ou de l'extraire à nouveau. On peut voir à la figure 7.12 au point #1, les jeux de données existant de l'utilisateur. Au point #2 (figure 7.12), l'utilisateur pourra consulter le détail de la structure du jeux de données tel que le nom des champs, les types de champs (numérique, texte, date).

Il sera possible de créer un nouveau jeu de données. Le point #3 (figure 7.12), permet de saisir le nom du jeu de données et d'y associer une description. Le point #4 présente le nom des concepts de l'ontologie des données à l'utilisateur. Lorsqu'un concept est sélectionné, ses champs associés s'affichent au point #5 et la définition du concept s'affiche au point #9. Si un champ présente une modification structurelle dans le temps, par exemple, sa taille est passée de 45 à 75, ce champ s'affichera au point #6 en rouge. L'utilisateur utilisera les flèches au point #7 pour déplacer les attributs qui feront partie de l'extraction. Lorsque l'on clique sur un champ au point #5, sa définition s'affiche aussi au point #9. Le point #8 indique l'ensemble des attributs finaux qui seront extraites dans un fichier plat. Le point #10 permet de naviguer dans le glossaire, le dictionnaire de données de l'entrepôt. Ce dictionnaire est complémentaire à l'ontologie des données et permet des liens à travers les mots de même sens.

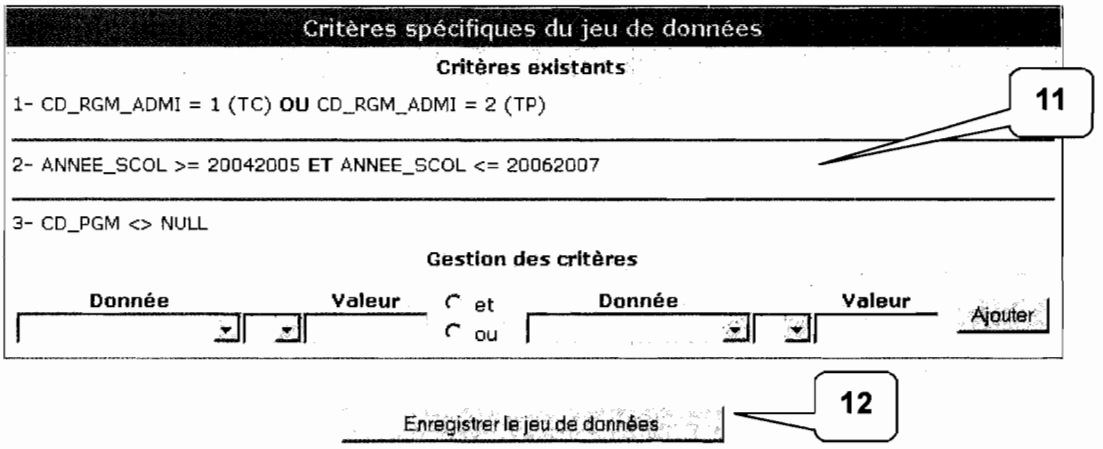


Figure 7.13 Interface d'extraction (API) – partie du bas.

Le point #11 de la figure 7.13 permet à l'utilisateur d'affiner son extraction. Il pourra créer des filtres permettant d'inclure ou d'exclure des données. Le point #12 permet l'enregistrement des configurations du jeu de données et la création physique du fichier plat.

Ce chapitre présente les résultats du travail de recherche. Il permet au lecteur de connaître les travaux qui seront amorcés dans la prochaine année. Il permet aussi d'aborder les travaux futurs connexes qu'ils auraient été intéressants d'étudier. Ce chapitre est divisé en deux points :

8.1 Résultats

8.2 Travaux futurs

8.1 Résultats

L'expérimentation du prototype a permis de mettre en exécution le guide méthodologique du chapitre 4 dont la synthèse reposait sur la littérature du chapitre 3. Le tableau 8.1 nous montre la synthèse de la méthodologie proposée.

Tableau 8.1

Synthèse de la méthodologie proposée

	Étapes	Choix retenus
1	**Approche de base**	***«bottom-up»*** (Kimball)
2	**Approche «orientée»**	**Hybride** (objectifs et besoins)
3	**Architecture logique**	**Dimensionnel *«bottom-up»* avec dimension et faits conformes** (avec *staging area non permanent*)
4	**Cycle de vie décisionnel**	Méthode Kimball
5	**Architecture physique**	**ROLAP**
6	**Modèle architectural des données**	**À 4 niveaux** (avec l'arbre du sujet[2])

Au chapitre 5 l'introduction du processus ETC nous a fait comprendre les étapes d'intégration des données sources vers l'entrepôt. Au chapitre 5, l'étude des outils existants nous a permis de faire une recommandation d'achat. L'implication de plusieurs ressources humaines de différentes compagnies et de différents agendas est assez complexe à gérer. Beaucoup de temps a été alloué à cette partie.

Résumons le processus d'évaluation. Une enquête préliminaire a permis de faire l'inventaire des produits sur le marché. Les logiciels libres «Open sources» autant que les progiciels corporatifs ont été étudiés. Cet inventaire a servi d'introduction à la première évaluation sur les fonctionnalités des suites logicielles. Voici les sept logiciels des six compagnies qui ont été évalués : JasperSuite de JasperSoft, Pentaho, SAS, Data Manager (Cognos), Cognos8 (Cognos), PowerCenter (Informatica) et Crystal Decisions de Business Objects (BO). De cette première évaluation, aucune recommandation n'a pu être retenue. Il a fallu faire une deuxième évaluation plus détaillée et plus technique en excluant l'outil PowerCenter d'Informatica car son prix pour son outil ETC était de $400 000.

Restait dans la course les cinq compagnies suivantes à évaluer : COGNOS, SAS, BO, JasperSoft et Pentaho. Pour ce faire, un plan technique des étapes à réaliser avec chaque logiciel fut établi. Ce plan a permis d'évaluer par pointage chacune des fonctionnalités. La compagnie COGNOS a obtenu une cote totale de 130 tandis que l'outil BO s'est gratifié d'une cote de 141. La suite BO fut légèrement supérieure. Nous n'avons pu évaluer SAS avec cette grille faute d'installation du logiciel. Suite à la réalisation de ce plan technique pour chaque compagnie, une deuxième évaluation plus détaillée a pu être réalisée. Cette dernière évaluation des outils en fonction des critères du tableau 5.6 a permis de recommandée la compagnie Business Object (BO) pour l'achat de la suite logiciel incluant un logiciel de chargement pour l'entrepôt de données de l'UQTR et un outil de présentation des données intuitif pour les utilisateurs externes. Les compagnies COGNOS, BO et SAS ont eu respectivement les pointages de 64%, 74% et 38% tandis que les compagnies JasperSoft et Pantaho ont eu respectivement 42% et 32%. Finalement, un devis technique pour l'achat du logiciel BO fut déposé au Service des achats de l'UQTR. L'installation de la suite est prévue à la session d'automne 2008. À la session d'hiver 2009, nous débuterons la création d'un premier processus d'affaires afin de produire notre premier tableau de bord

pour la fin de la session d'hiver 2009. Le tableau 8.1 nous présente le résumé de l'étude des outils existants.

Tableau 8.2
Résumé de l'étude des outils existants

Enquête préliminaire	Inventaire des outils
Évaluation #1	Études des outils sélectionnés à l'enquête préliminaire : aucune solution retenue.
Plan technique	La suite BO est légèrement supérieure.
Évaluation #2	Évaluation détaillée des outils en fonction des critères du tableau 5.6 : La suite BO est supérieure.
Devis (recommandation d'achat)	La suite BO est recommandée pour l'achat

Au chapitre 6, le prototype d'entrepôt de données fut réalisé. Dans la phase d'analyse des données, deux axes ont été explorés. L'axe de profondeur des données d'une table et l'axe relationnel des données entre deux tables. Cette analyse fut assez longue et minutieuse. Elle a permis la création d'outils d'analyse qui permettront d'avoir une rigueur au niveau du développement des systèmes transactionnels sources. Nous avons créé le logiciel SAT (Système d'analyse de tables) qui permet l'analyse de fréquence des valeurs d'une table et l'analyse des relations.

La mise en place de clés étrangères, dès le début d'un projet OLTP, permettra d'augmenter la rigueur d'un cran. Ces clés étrangères facilitent la dénormalisation d'une relation entité-relation en dimension. Pour ce faire, on passe de la 3^{e}NF à la 2^{e}NF. Découlant du chapitre 6, la phase de préparation des données est l'une des phases les plus importantes. Nous proposons qu'en premier lieu l'analyse unitaire des tables du schéma ER du besoin à implanter soit faite. Par la suite, une analyse de chaque relation devra indiquer l'état de santé de celle-ci. S'il y a lieu, placer les clés étrangères requises du côté OLTP et ce, même si cette proposition peut avoir des conséquences majeures de remaniement des OLTP.

À la fin du chapitre 6, les données étaient intégrées à l'entrepôt par des programmes développés en PL/SQL au SSPT. Puisque le choix du produit ETC n'avait pas été établi, les huit phases du développement du processus, ETC pour l'élaboration d'un besoin devront être évaluées avec la suite BO. Deux méthodes de chargement ont été testées, le chargement avec les vues matérialisées et le chargement avec le CDC d'Oracle. Le chargement avec le CDC d'Oracle sera choisi et servira de zone de chargement temporaire des données pour la suite BO en amont de son outil ETC.

Nous avons créé deux autres logiciels pour les gestionnaires de l'entrepôt. Le premier, le logiciel OAD (Outils d'analyse des DDL) permettant d'envoyer au gestionnaire de l'entrepôt toutes les modifications des structures des tables des systèmes transactionnels. Le gestionnaire, en consultant la liste des modifications, pourra décider ou non de modifier les processus de chargement de l'entrepôt. Puisqu'à l'UQTR, le gestionnaire de l'entrepôt n'est pas la même personne, il serait insensé de demander au DBA d'aviser le gestionnaire de l'entrepôt à chaque modification du côté de la base de données. Le logiciel OAD permet ce suivi.

Par la suite, le logiciel SEPTS (SystèmE de Perception Temporelle des Structures) fut créé et est encore en développement. Ce logiciel permet l'historisation des structures de l'entrepôt de données. Tout comme l'historique des données de l'entrepôt, toutes les modifications aux structures des tables du côté de l'entrepôt seront aussi emmagasinées dans le temps. Cette idée a permis d'entrevoir la possibilité d'effectuer cet historique des structures du côté des systèmes transactionnels. Une adaptation majeure au logiciel SEPTS est en cours. Les analystes des systèmes transactionnels vont passer par ce système pour «commander» aux DBA leurs créations de table, leurs modifications ou leurs suppressions. Les DBA, vont exécuter ou modifier ces demandes et une fois acceptées, l'historique des modifications sera sauvegardée. On pourra alors consulter la liste des modifications d'une table de sa création à aujourd'hui ou à un moment précis dans le temps.

Au chapitre 7, la publication et la présentation des données ont été introduites. Afin de rendre accessibles les données au gestionnaire, l'outil de présentation «Report studio» de COGNOS a été testé. Préalablement à l'outil «Report Studio», la préparation du moteur

ROLAP de COGNOS l'outil «Framework manager» fut configuré. Il fallait charger les structures dimensionnelles que l'on voulait rendre accessibles aux utilisateurs dans le modèle physique. On utilisait le modèle logique des données pour la fabrication des tableaux de bord par l'usager. Les hiérarchies de forage des données fut créée. Une fois les modèles publiés, l'utilisateur avait accès aux structures dimensionnelles dans le portail COGNOS8 afin de construire ses rapports et ses tableaux de bord. Trois rapports de bord ont été créés. Le premier, la liste des étudiants admis par programme selon le régime d'étude à temps complet (TC) ou à temps partiel (TP). Le second étant une autre vue sur les mêmes données soit la liste des étudiants admis par départements en fonction de leur régime d'étude. Le dernier, la liste des étudiants admis par cycle selon le régime d'étude. Ces rapports créés ne sont pas nécessairement des tableaux de bord au sens propre puisqu'ils n'ont pas d'indicateur graphique permettant de prendre visuellement et rapidement des décisions. Faute de temps, les fonctions d'indicateurs n'ont pas été testées. On aurait pu mettre en rouge les demandes d'admission sous le seuil de viabilité d'un programme ou d'un département ce qui impliquait d'obtenir l'information «seuil de viabilité» d'un département et donc le processus financier d'un département qui est très complexe.

Finalement, à la section 7.3, un prototype d'interface d'extraction des données fut proposé. Cette proposition est un complément qui permettra à d'autres logiciels de recevoir les données de l'entrepôt à des fins d'analyse plus fine telle que le *data mining*. Ce prototype ne s'est pas rendu à la phase de programmation. Il aurait été intéressant de pouvoir importer des fichiers plats de l'entrepôt de données de l'UQTR vers le logiciel «WEKA» pour des analyses de découverte de patron, de connaissance.

8.2 Travaux futurs

Au cours de la prochaine année, nous prévoyons procéder à l'intégration de nouveaux processus d'affaires qui se grefferont à l'entrepôt de données. Nous pourrons ainsi vérifier, améliorer et valider la méthodologie proposée. La mise en application des objectifs du projet et des suites à y apporter sera effectuée dans le cadre du travail régulier de l'auteure du présent mémoire. En effet, embauchée à titre de professionnelle, comme analyste de l'informatique au Service de soutien pédagogique et technologique de l'Université, l'auteure ainsi que son collègue Michel Charest ont été mandatés pour assurer le suivi de l'implantation du projet dans son ensemble. Ainsi, l'Institution s'assure d'obtenir une grande cohérence dans la planification et l'application des mesures proposées tout en ayant l'auteure de ces propositions sur place en permanence.

Le projet est déjà bien en place même que le logiciel SEPTS sera implanté dans quelques semaines mais seulement du coté de l'entrepôt de données. Les outils SAT et OAD, quant à eux, seront modifiés afin d'améliorer certaines fonctionnalités.

L'achat de l'outil ETC et de l'outil de présentation se concrétisera afin d'automatiser l'ensemble des fonctionnalités de l'entrepôt de données. Dans ce sens, certains groupes d'utilisateurs seront sollicités afin de participer à la réalisation et à la mise en place de leurs tableaux de bord personnalisés. Le bureau de la réussite étudiante (BRE) de l'UQTR sera le premier service qui bénéficiera des services de l'entrepôt institutionnel de données.

En considérant le temps alloué à la section de l'étude des outils existants et celle de la réalisation du prototype de l'entrepôt de données; des choix se sont imposés quant à la limitation du projet. Il aurait été très intéressant de développer davantage les fonctionnalités des tableaux de bord en utilisant des indicateurs graphiques ou encore de réaliser un nouveau processus d'affaires du début à la fin afin de vérifier notre méthodologie. L'installation et l'utilisation de la suite logicielle de BO auraient permis de conclure le projet avec une présentation concrète aux dirigeants.

Il existe des travaux qu'ils auraient été intéressants d'étudier et d'inclure au projet. Il s'agit de l'implantation de l'ontologie des données de l'entrepôt qui défini les concepts et le vocabulaire unique du domaine. Ces concepts et définitions pourraient être accessibles pour l'utilisateur. L'autre projet, dans le contexte de l'UQTR, permettrait d'effectuer des prédictions sur les données de l'entrepôt. On utiliserait un système de raisonnement à base de cas, branché à l'entrepôt, afin de prédire la réussite étudiante en comparant les données d'un étudiant au patron établi «RÉUSSITE». Il convient d'introduire ces deux sujets.

Ontologie des données

Une ontologie est un ensemble de termes hiérarchiquement structurés dérivant les concepts généraux relatifs à un domaine. Il fournit le vocabulaire et les relations entre les concepts. Cette définition est très liée au dictionnaire de données de l'entrepôt. L'ontologie permettrait de définir une seule fois un concept. Le concept ainsi défini est donc univoque peu importe l'utilisateur.

Chaque table, chaque donnée de l'entrepôt sera définie dans la structure ontologique. Il serait intéressant d'alimenter automatiquement l'ontologie lorsque l'on ajoute ou modifie une donnée dans l'entrepôt et ce, en utilisant une interface «API» permettant d'effectuer ce genre d'opération. La définition des données de l'ontologie, intégrée au glossaire du système, informerait l'utilisateur, notamment en ce qui concerne le vocabulaire du domaine universitaire.

La figure 8.1 permet d'afficher un glossaire avec la définition des concepts. Chaque concept peut être éclaté pour obtenir des informations supplémentaires sur la composition des éléments du concept choisi. De plus, pour combler une notion importante des métadonnées, un lien vers les éléments de même sens pourrait être accessible en offrant à l'utilisateur la possibilité de naviguer à travers la toile hiérarchique de l'information.

Concept	Description
Admission	L'admission comprend l'ensemble des activités requises pour examiner les candidatures d'étudiants à des programmes de 1er cycle et de cycles supérieurs, contingentés ou non. Le processus débute par la réception des demandes et se termine par la décision d'admission en passant, s'il y a lieu, par la gestion des tests et des entrevues.
Étudiant	Un étudiant est une personne admise à l'Université en vue d'une formation. Voir l'article 1.1 du Règlement no. 5

Caractéristiques	Description	Synonyme
Code permanent	Code d'identification unique d'un étudiant	Code d'identification
Nom	Nom légal de l'étudiant au ministère de l'éducation	
Prénom	Prénom légal de l'étudiant au ministère de l'éducation	
...		

Inscription	Opération par laquelle l'étudiant régulier procède à son choix de cours, selon les dates prévues à cet effet, parmi les cours qui composent son programme d'études et qui sont offerts au trimestre concerné.
Programme	Un programme est un ensemble cohérent de cours portant sur une ou plusieurs disciplines, sur un ou plusieurs champs d'études ordonné à une formation définie par son principe intégrateur. Voir l'article 2.2 du Règlement no. 5

Caractéristiques	Description	Synonyme
Code de programme	Code d'identification unique d'un programme	Code d'identification
Nom	Nom du programme	Nom abrégé 1 Nom abrégé 2
...		

Figure 8.1 L'ontologie des données en ligne de l'entrepôt de données.

La mémoire de maîtrise de ma collègue Yanfen Shen du groupe de recherche en *data mining* [SHEN 07] fourni les explications essentielles à l'étude et à la mise en place de l'ontologie des données.

Data mining et RBC (Raisonnement à base de cas)

La métaphore du *data mining* signifie qu'il y a des trésors ou pépites cachés sous la montagne de données que l'on peut découvrir avec des outils spécialisés. C'est une combinaison d'outils informatiques (base de données), d'intelligence artificielle (apprentissage automatique) et de statistiques permettant d'analyser les informations existantes pour obtenir d'autres informations qui sont une base de connaissance utile pour les décisions stratégiques et opérationnelles.

Dans le contexte de l'UQTR, on pourrait prédire la réussite étudiante en comparant les données d'un étudiant au patron établi «RÉUSSITE». L'agent intelligent nous indiquerait alors si un étudiant est en voie de réussite ou non, ainsi que la probabilité de ce fait.

Il serait intéressant d'intégrer l'agent à l'entrepôt de données et à l'aide d'un outil RBC (Raisonnement à base de cas) d'aider l'utilisateur dans ses choix pour la construction de ses patrons. Les patrons et les résultats pourront être ainsi emmagasinés afin d'inciter l'utilisateur à la réutilisation. La figure 8.2 illustre le schéma de l'intégration de l'agent, du *data mining* et de l'ontologie à l'entrepôt.

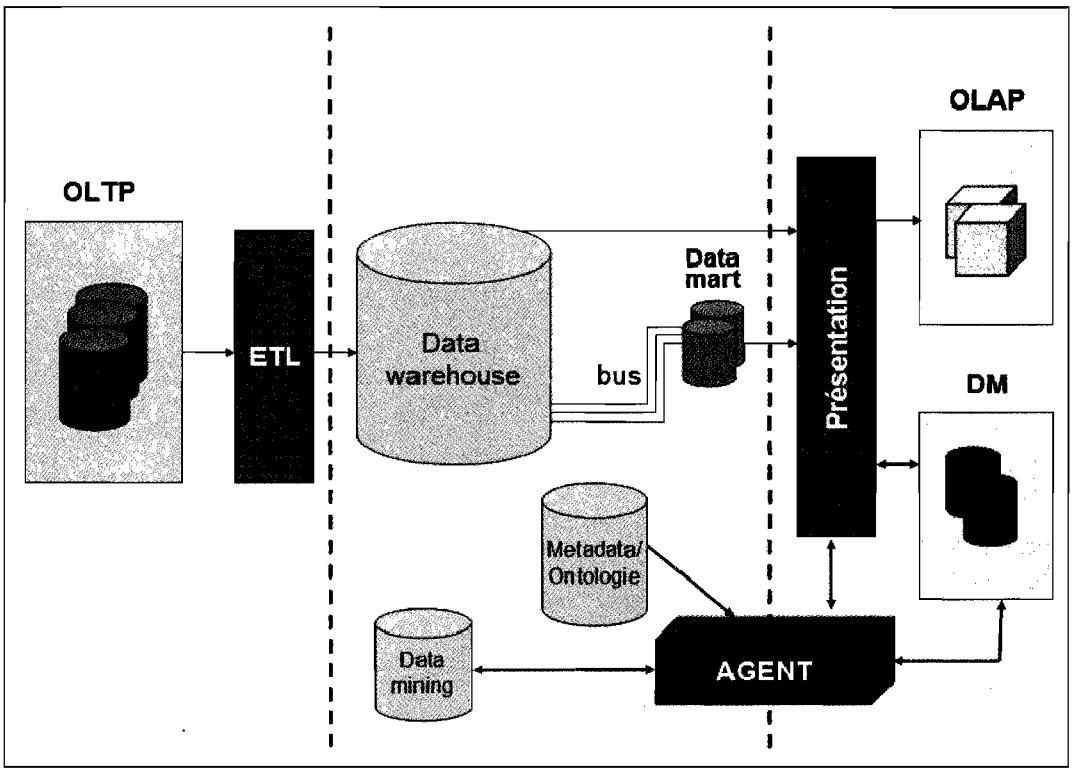


Figure 8.2 L'agent intelligent RBC pour l'aide au *data mining*.

La mémoire de maîtrise de mon collègue Michel Charest, du groupe de recherche en *data mining* [CHAREST 07] et collègue au SSPT, fourni les explications essentielles à l'étude et à la mise en place d'un agent RBC à l'entrepôt.

Chapitre 9
Conclusion

Le présent travail de recherche veut fondamentalement faciliter la prise de décision devant les nombreuses orientations que peuvent prendre le Service de soutien pédagogique et technologique ainsi que les gestionnaires de l'Université.

Nous sommes conscients que la lecture de notre document peut paraître parfois complexe pour le lecteur. Il était cependant indispensable pour nous, afin de bien faire comprendre le développement proposé et l'évolution des systèmes, d'utiliser le vocabulaire spécialisé et approprié à notre étude.

Malgré les aspects techniques et la terminologie, parfois hermétique, de notre rapport de recherche, celui-ci constitue une première étape qui fait appel au besoin de la productivité et de l'efficacité maintenant exigées par les utilisateurs des systèmes informatisés. Notre recherche, ne le cachons pas, n'est pas neutre dans le sens où l'auteure fait face à des choix qui sont déterminants en termes de conséquences à court et à long termes.

Dans ce document, nous venons de proposer une méthodologie servant de guide à la réalisation d'un entrepôt de données. Cette phase d'exploration, de découverte et d'appropriation du domaine fut une phase initiale nécessaire et essentielle à la mise en place du prototype et à la poursuite du projet dans le cadre de notre travail. Elle permet de décrire toutes les étapes de création et d'exploitation d'un entrepôt de données.

Pour se remettre dans le contexte initial, résumons brièvement les objectifs de ce document.

- Créer un prototype d'entrepôt institutionnel de données et proposer une méthodologie de conception.
- Guider les dirigeants à schématiser leurs processus d'affaires.
- S'assurer de l'intégrité référentielle des systèmes sources.
- S'assurer de la fiabilité et de l'exactitude des résultats en informant les dirigeants.

Le premier objectif était de créer un prototype d'entrepôt institutionnel de données et proposer une méthodologie de conception. L'étude des différentes approches de base nous a permis de faire un choix sur lequel reposera l'entrepôt de données de l'UQTR. L'objectif du mémoire découle d'un besoin de support d'aide à la prise de décision. La réalisation du prototype du suivi de l'admission a permis de constater la facilité avec laquelle un tableau de bord peut être construit lorsque l'on a accès aux structures de l'entrepôt de données. Ce premier objectif est atteint dans la mesure ou chacune des étapes de la création du prototype de l'entrepôt fut tester et implanter. Cependant, seul l'installation de la suite logicielle de BO aurait permis une connexion en temps réelle entre partie de l'architecture du l'entrepôt soit les données sources, le CDC, l'ETC, le moteur ROLAP et l'outil de présentation des données.

Le second objectif était d'aider les dirigeants à exprimer et à schématiser leurs processus d'affaires. Le questionnaire destiné aux dirigeants permet de mettre en relation la vision des processus d'affaires des demandeurs en les arrimant aux objectifs plus globaux de leur service. En les aidant à schématiser tout en basant les processus besoins vers les objectifs globaux, on limite les changements radicaux de leurs besoins et par conséquent, les modifications associées à ces besoins dans l'entrepôt de données. Cet objectif est atteint mais la participation de plusieurs utilisateurs permettrait d'affiner le questionnaire et d'améliorer son analyse.

Le troisième objectif consiste à combler la lacune au niveau de la qualité et de l'intégrité des données dans les systèmes servant de source à l'entrepôt de données. Malgré la multitude d'articles sur le sujet, nous avons été surpris de constater qu'aucune documentation par rapport à l'analyse fine des données sources n'avait été produite. Cependant, on peut lire maintes et maintes fois que la phase de préparation des données est la phase qui requiert plus de 75% du temps de réalisation d'un projet et qu'il ne faut pas la négliger. Nous pouvons conclure que la qualité du développement des systèmes sources est un facteur déterminant à la réalisation d'un entrepôt de données fiables. Nous avons mis davantage l'emphase sur la préparation des données plutôt que sur leur chargement. Ce troisième objectif est atteint par la création d'outils logiciels soit les outils SAT et OAD. Ces outils d'analyse des structures ont été réalisés afin d'affiner l'analyse de la qualité des données et de faciliter la gestion de la phase de prétraitement. On peut

vérifier l'intégrité avec le logiciel SAT et ainsi déterminer à quel endroit corriger le problème.

Finalement, afin de combler le dernier objectif, le logiciel SEPTS fut créé. Ce logiciel permet de créer l'historique des structures. Cet historique combiné avec les informations des événements journaliers de la vie à l'UQTR dans la table de dimension temps (DIM_TEMPS) permet d'informer un gestionnaire d'un événement ou d'une modification de table qui influencerait le résultat de ses extractions selon sa période choisie. Cet objectif est atteint partiellement, car il n'y a aucun moyen d'interagir avec le logiciel de présentation des données de BO. Par contre, en sachant que l'on possède cette richesse informationnelle, nous aurons le loisir de l'interroger parallèlement.

Au lieu d'avoir utilisé un outil ETC pour le chargement de l'entrepôt, des programmes PL/SQL ont permis de charger les données. Ces programmes pouvaient charger les données selon une certaine fréquence ou exécuter un chargement initial des données. L'achat d'un outil de chargement (ETC) permettra d'utiliser les huit phases du développement ETC, de les implanter et de les tester. L'installation de la suite logicielle de BO est à réaliser afin de pouvoir poursuivre le projet.

La méthodologie résultante se veut un guide de base pour éclairer toute personne désirant amorcer un projet d'entrepôt de données. La méthodologie proposée devient le cadre qui orientera de façon plus éclairée les chemins les moins fréquentés. Il faut voir les données d'un autre œil, d'une autre perspective. L'ampleur du projet et de ses réalisations pour résoudre les problèmes rencontrés dénote l'accomplissement du travail et des objectifs du projet.

Une proposition d'un outil d'extraction de données qui intègre l'analyse des impacts temporels et événementiels permettra d'informer le demandeur des biais possibles dans l'extraction des données. Le fichier résultant de l'extraction demandée pourra éventuellement devenir la source d'un logiciel de *data mining*.

En terminant, précisons que nous ne prétendons pas avoir réponse à toutes les questions sur le développement de nos systèmes mais disons, humblement, que notre étude ouvre grande la porte à la planification prospective dans le domaine de l'informatisation des données.

Le présent ouvrage fournit donc un cadre de référence qui permettra à l'Université de procéder à l'évolution des systèmes dans une perspective de développement optimal.

LISTE DES RÉFÉRENCES

[AMO-ALVES 00] AMO, S., Halfeld Ferrari Alves, M., 2000. « *X-META: A Methodology for Data Warehouse Design with Metadata Management* ». IEEE International Conference on E-Commerce Technology.

[CARNEIRO-BRAYNER 00] Carneiro, L., Brayner, A., 2000. « *Representing Temporal Data in Non-Temporal OLAP Systems* ».

[EDER-KONCILIA 02] Eder, J., Koncilia, C., 2002. « *Using AutoMed Metadata in Data Warehousing environments* ». IEEE International Conference on E-Commerce Technology.

[ECKERSON 03] Eckerson, W., 2003. « *Four ways to build a data warehouse* ». TDWI, volume 15, may 2003

[FAN-POULOVASSILLIS 03] Fan, H., Poulovassillis, A., 2003. « *Building the Data Warehouse* » .PORTAL : The ACM digital library.

[GARDNER 98] Gardner S.. 1998. « *Goal-Oriented Requirement Analysis for Data Warehouse Design* ». PORTAL : The ACM digital library.

[GIORGINI et al. 05] Giorgini, P., Rizzi, S., Garzetti, M., 2005. « *A Methodological Framework for Data Warehouse Design* ». PORTAL : The ACM digital library.

[GOLFARELLI-RIZZI 98] Golfarelli, M., Rizzi, S., 1998. « *The DSS environment DW-Date marts and data mining* ». PORTAL : The ACM digital library.

[INMON 02] Inmon, W. H., 2002. « *Modeling Strategies and Alternatives for Data Warehousing Projets* ».

[JURIC 06] Juric, N., 2006. « *Best Practices in Data Warehouse to Support Business Initiatives and Needs* ». PORTAL : The ACM digital library.

[LAWER-CHOWDHURY 07] Lawer, J., Chowdhury, S., 2007. « *Report on the 5th International Workshop on the Design ans Management of Data Warehouse (DMDW'03)* ». IEEE International Conference on E-Commerce Technology.

[LENZ et al. 03] Lenz, H., Vasiliadis, P., Jeusfeld, M., Staudt, M., 2003. « *A Comparison of Data Warehouse Development Methodologies* ». PORTAL : The ACM digital library.

[LIST et al. 02] List, B., Bruckner, R., Machaczek, K., Schiefer, J., 2002. « *Strategy and Approach for the Next-Generation Data Warehouse* ». DEXA 2002.

[PEIPERT-ALBALA 05] Peipert, G., Albala, M., 2005. « *A Comparison of Data Warehousing Methodologies* ». Conversion Services International, Inc..

[SEN-SINHA 05] Sen, A., Sinha, A.P., 2005. « *Exploiting bitemporal schema versions for managing an historical medical data warehouse: A case study* ». PORTAL : The ACM digital library.

[SERNA-ADIDA 05] Serna-Encinas, M.-T., Adiba, M., 2005. « *Data Warehousing With Oracle* ». IEEE International Conference on E-Commerce Technology.

[SHAHZAD 00] Shahzad, M., 2000. « *Power System Data Warehouses* ». Oracular.

[SHI et al. 01] Shi, D., Lee, Y., Duan, X., Wu, Q.H., 2001. « *The Four-stage Standardized Modeling in Data Warehouse System Development* ». IEEE International Conference on E-Commerce Technology.

[SHUNUNG et al. 05] Shunung, W., Deguang, C., Peng, C., 2005. « *DATA WAREHOUSE PROCESS MANAGEMENT* ». IEEE International Conference on E-Commerce Technology.

[TDWI 04] 2004. « *Efficeint Maintenance of Temporal Data Warehouses* ». TDWI.

[VASSILIADIS et al. 01] Vassiliadis, P., Quix, C., Vassiliou, Y., Jarke2, M., 2001. « *A Method for Demand-driven Infromation Requirements Analysis in Data Warehousing Projets* ».

[WINTER-STRAUCH 02] Winter, R., Strauch B., 2002. « *Multiversion Data Warehouses : Challenges and Solutions* ». PORTAL : The ACM digital library.

[WREMBEL-MORZY 05] Wrembel, R., Morzy, T., 2005. IEEE International Conference on E-Commerce Technology.

BIBLIOGRAPHIE

Mémoires :

[CHAREST 07] Charest, M., 2007. «*Intelligent Data Mining Assistance via Case-Based Reasoning and a Formal Ontology*». Mémoire de maîtrise, Université du Québec à Trois-Rivières.

[DUGRE 04] Dugré, M., 2004. « *Conception et réalisation d'un entrepôt de données.*». Mémoire de maîtrise, Université du Québec à Trois-Rivières.

[SHEN 07] Shen, Y., 2007. «*A Formal Ontology for Data Mining: Principles, Design, and Evolution*». Mémoire de maîtrise, Université du Québec à Trois-Rivières.

Webographie :

Entrepôt de données
http://www.rkimball.com/
http://www.datawarehouse.org/
http://www.datawarehousing.com/
http://www.inmoncif.com/home/
www.tdwi.org

Data Mining
www.web-datamining.net
http://nakache.9online.fr/probatoire/
http://eric.univ-lyon2.fr/~ricco/data-mining/

Tableau de bord
http://www.nodesway.com/
http://www.le-perfologue.net/

ETC
http://www.systemeETC.com/

OLAP
Les bases de données OLAP
OLAP Report

Information
http://www.lemondeinformatique.fr/
http://www.journaldunet.com/solutions/dossiers/pratique/entrepot-donnees.shtml

Modélisation conceptuelle d'une université selon Inmon
http://inmoncif.com/registration/datamodels/models/univrsty/univer.php

Livres :

[Fernandez 05] Fernandez, A. (2005). *L'essentiel du tableau de bord.* Editions d'Organisation, ISBN : 978-2-7081-3104-0

[Inmon 96] Inmon, W. (1996). *Building the Data Warehouse*, Second Edition, John Wiley and Sons.

[Voyer 02] Voyer, P. (2002). *Tableaux de bord de gestion et indicateurs de performamce.* Presses de l'Université du Québec, 446 pages, ISBN : 2-7605-0991-5

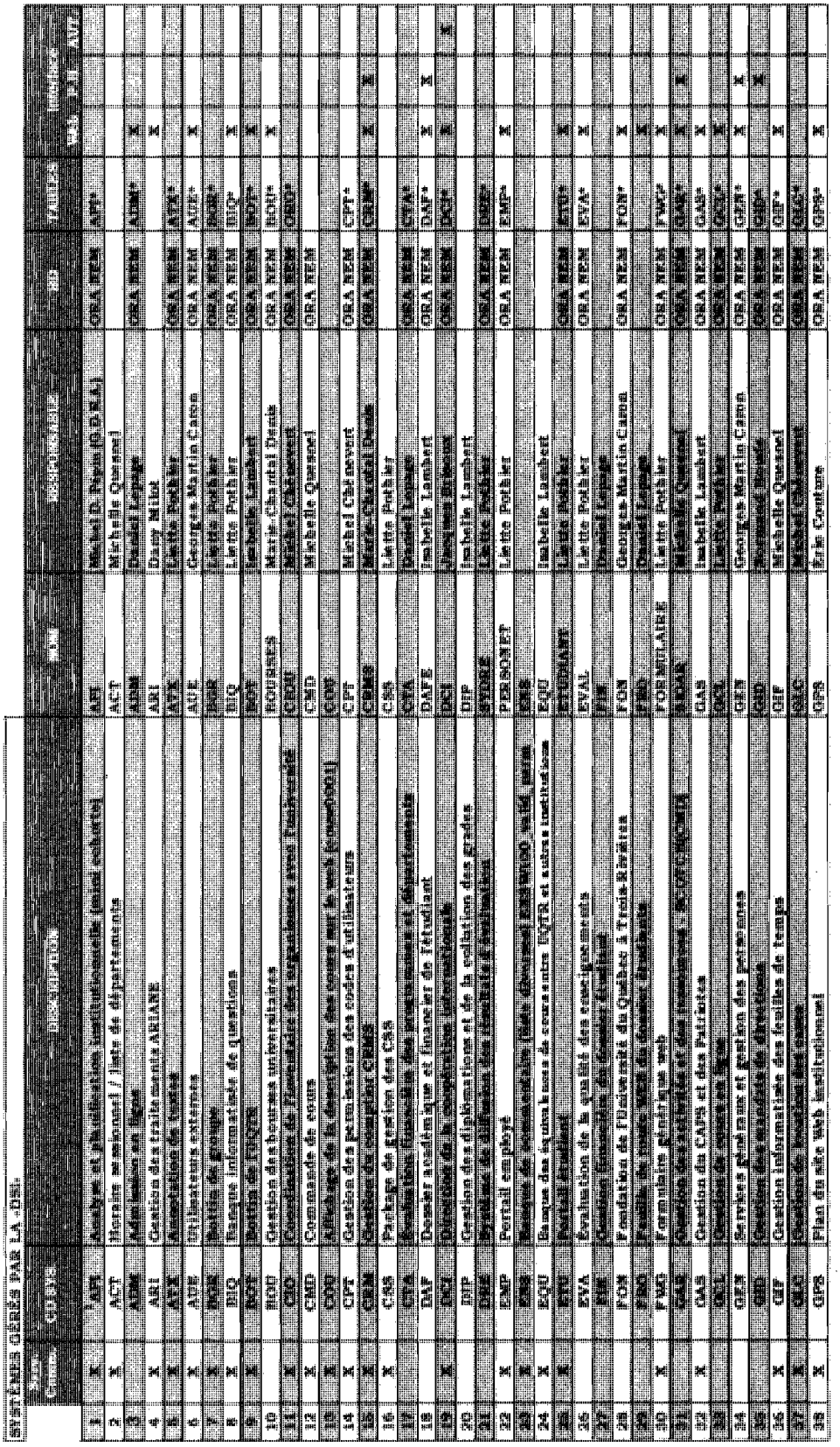
[Kimball et al. 05] Kimball, R., Reeves, L., Ross, M., Thornthwaite, W. (2005). *Le data warehouse Guide de conduite de projet.* EYROLLES France, 576 p., ISBN : 2-212-116004

[Kimball et al 04] Kimball R, Caserta. (2004). *The Data Warehouse ETC Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. John Wiley & Sons, 2004 (416 pages).

ANNEXES

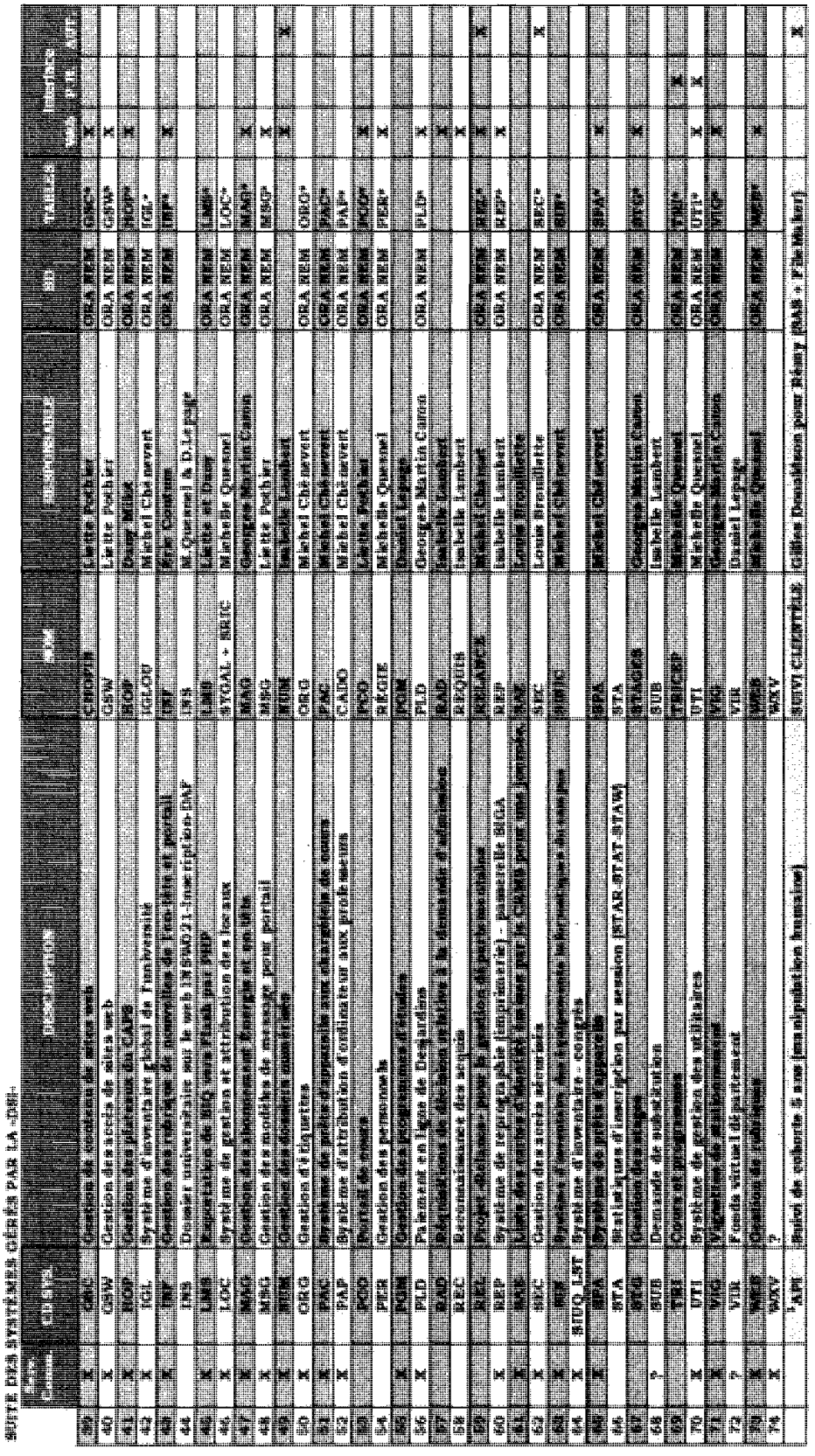
ANNEXE A

«Inventaire des systèmes de l'UQTR»



ANNEXE A

«Inventaire des systèmes de l'UQTR»



ANNEXE A

«Inventaire des systèmes de l'UQTR»

AUTRES SYSTÈMES DE L'UQTR

	Serv. Comm.	CD SYS.	DESCRIPTION	NOM	RESPONSABLE	BD	TABLES	Interface		
								Web	P.H.	AUT.
75		SIGA	Système de gestion administratif : FINANCES	SIGA	Serveur + passerelle oracle					X

EXTERNE

	Serv. Comm.	CD SYS.	DESCRIPTION	NOM	RESPONSABLE	BD	TABLES	Interface		
								Web	P.H.	AUT.
76		ICOPE	Indicateurs de COnditions de Poursuite des Études	ICOPE	Rémy Auclair	ORA NEP	ICO			X
77	X	CREPUQ	Organisme externe : Programme d'échange étudiants Demande de cours en commandite dans une autre institution Transmission de la mise à jour des programmes	CREPUQ	Réseau UQ (Registraire d'accueil accepte ou non)					X

ORA NEM Base de données oracle sur serveur «Némésis»
ORA NEP Base de données oracle sur serveur «Neptune»

SAS Logiciel SAS pour statistique
SIGA Logiciel sur serveur NEMESIS avec SID Oracle «SIGASS SIGAPP»
Filemaker Logiciel utilisé pour l'impression du rapport du suivi sur 5 ans

Web Programme converti en page web transactionnelle
P.H. Programme dont les interfaces sont définie sous Power House (COGNOS)
Aut. Exclus les deux autres (Web et P.H.)

ANNEXE B

«Méthode en 5 étapes et 14 outils pour l'élaboration d'un tableau de bord»

Etape 1 Sélectionner les axes d'action	- Identifier les principales sources de revenu - Situer l'entreprise sur son marché - Evaluer les attentes des clients - Repérer les principaux leviers - Evaluer et choisir les axes de progrès	Outil n°1 Outil n°2 Outil n°3 Outil n°4 Outil n°5
Etape 2 Déterminer les points d'intervention	- Identifier les processus et les activités critiques	Outil n°6
Etape 3 Sélectionner les objectifs	- Choisir les objectifs - Mesurer les risques - Elaborer les plans d'action	Outil n°7 Outil n°8 Outil n°9
Etape 4 Sélectionner les indicateurs	- Choisir les indicateurs - Présenter l'indicateur sur le poste de travail	Outil n°10 Outil n°11
Etape 5 Structurer le tableau de bord	- Construire le tableau de bord (vue de signalisation) - Organiser le tableau de bord (vue d'analyse et de prospection) - Maintenir le tableau de bord	Outil n°12 Outil n°13 Outil n°14

Tableau extrait du livre de Fernandez [FERNANDEZ 05]

Voici en résumé les étapes de réalisation des tableaux de bord. Dans son libre, l'auteur explique en détail les étapes et leurs outils correspondants.

ANNEXE C

«Classement des systèmes OLAP»

Le tableau suivant permet de classifier chaque type d'architectures et les fournisseurs de solutions OLAP selon deux critères : La technologie de stockage et le traitement OLAP.

Source : www.systemeETC.com		Technologies de stockage de données MULTIdimensionnelle		
		Base de données relationnelle	**Base de données dimensionnelle**	**Fichier sur le poste client**
Traitement OLAP	**SQL Multiples passes**	ROLAP Cartesis Magnitude MicroStrategy		
	Serveur de traitement OLAP	ROLAP,HOLAP Crystal Holos (ROLAP mode) Hyperion Essbase Longview Khalix Speedware Media/MR Microsoft Analysis Services Oracle Express (ROLAP mode) Oracle OLAP Option (ROLAP mode) Pilot Analysis Server WhiteLight	MOLAP, HOLAP SAS CFO Vision Crystal Holos Geac MPC Hyperion Essbase Oracle Express Oracle OLAP Option AW Microsoft Analysis Services PowerPlay Enterprise Server Pilot Analysis Server Applix TM1	
	Client de traitement OLAP	ROLAP Oracle Discoverer	ROLAP Comshare FDC Dimensional Insight Hyperion Enterprise Hyperion Pillar	ROLAP Hyperion Intelligence Business Objects Cognos PowerPlay Personal Express TM1 Perspectives

ANNEXE D

«Comparaison des infrastructures»

Attributes	NCR/Teradata Methodology	Oracle Methodology	IBM DB2 Methodology	Sybase Methodology	Microsoft SQL Server Methodology
Core Competency	Teradata DBMS (massively parallel DBMS)	Oracle DBMS	DB2 DBMS	Sybase DBMS	SQL Server DBMS
Requirements Modeling	Interview, JAD, Prioritization, templates, document analysis	Interview, Prioritization, subject areas	Interview, JAD	Interview	Interview document analysis
Data Modeling	ERD, relational schema	Dimensional model, Star schema	Dimensional model, Star schema	ERD, Star schema, Relational schema	Dimensional model, Star and Snowflake schema
Support for Normalization/ Denormalization	Develops all relations at normalized, allows denormalization	Allows both	Allows both	More geared toward denormalization	Allows both
Architecture Design Philosophy	Enterprise data warehouse	Data marts	Enterprise data warehouse and data marts	Data marts	Enterprise data warehouse and data marts
Implementation Strategy	Iterative	Dimensional Life Cycle	Iterative (prototyping)	Iterative (RAD)	Iterative
Metadata Management	Yes, uses a Repository	Yes, uses Oracle Repository	Yes, uses a Repository	Yes, uses a Repository	Yes, uses Microsoft Repository
Query Design	Parallel query development	Allows parallel query	Not reported	Not reported	Allows parallel query
Scalability	Yes, to hundreds of Terabytes	Not reported	Yes	Not reported	Yes, to Terabytes
Change Management	Has pros and cons, but not emphasized in the methodology	Not reported	Not reported	Has extension in the methodology	Not reported

Source : [SEN-SINHA 05]

ANNEXE D

«Comparaison des infrastructures»

Attributes	SAS Methodology	Informatica's Velocity Methodology	Computer Associates' Methodology	Visible Technologies' Methodology	Hyperion STAR Methodology
Core Competency	Data analysis	Data analysis	Business Intelligence and Middleware	Business analysis software	Business analysis software and OLAP server
Requirements Modeling	Interview,JAD,document analysis	Business process inventory,JAD, subject area	Interview,JAD, document analysis	Interview,JAD, presentation, templates,document analysis	Analyze data sources and data sources
Data Modeling	ERD,Dimensional model,Relational schema	ERD,Dimensional model,Star schema	ERD,Dimensional model,Star schema	Warehouse model, ERD,Star schema	Dimensional model, Star schema
Support for Normalization/ Denormalization	Not reported	Not reported	Not reported	Allows both	Allows both
Architecture Design Philosophy	Enterprise data warehouse and data marts	Enterprise data warehouse with data marts	Enterprise data warehouse with data marts	Enterprise data warehouse with data marts	Enterprise data warehouse with data marts
Implementation Strategy	Iterative	Iterative (rapid prototype)	Iterative (rapid prototype)	Iterative	Iterative
Metadata Management	Yes. Uses integrated metadata management	Yes. Uses an integrated metadata platform	Yes. Uses its own repository	Yes. Uses its own repository	Yes
Query Design	Depends on the DBMS to be used at the warehouse level	Allows parallelism	Not reported	Not reported	Allows parallelism via partitioning
Scalability	Yes	Yes	Yes	Yes	Yes
Change Management	Very little	Very little	Yes	Uses Visible tools	Not reported

Source : [SEN-SINHA 05]

BI/DW Software Era	EIS/DSS	BI/DW	Next Generation
Main focus	Insight for a small, self-contained audience; functionality	Single version of the truth; data organization and performance	Context-driven insight driving a single version of the truth across trading partners; data organization, functionality, and security
Audience	Senior management	Internal stakeholders	All stakeholders, customers, suppliers
Stated business drivers	Drowning in paper	Drowning in data	Data is too latent for the decision-making process
Typical data volume	<1 gigabyte	<5 terabytes	>1 exabyte
Data organization	Simple	Star schema or MDDB	Hybrid relational/object/model
Data quality	Quality issues surfaced through EIS & DSS efforts	Data quality must be baked in at the start of efforts	Data quality across participating stakeholders and business units mandatory
Data quality approach	Passive	Reactive	Proactive
Early signals requiring attention	Senior management buried in paper	Stovepipes or islands of BI applications	Disjointed portals, BPM, BAM, predictive modeling, BI/DW applications, & operational systems
Data refresh rate	Monthly	Daily	Real time
Data provisioning	Simple scripted process	ETL	Merged ETL/EAI/EII/BPM/ change data capture
Business intelligence approach	Reporting/OLAP	Scorecards/dashboards	Predictive analysis/ alerts & notifications
Educational vehicle	Specialized development team	Centers of excellence	To be evolved

ANNEXE F

«Questionnaire aux dirigeants»

Partie 1 : Identification

1.1- Nom :

1.2- Prénom :

1.3- Poste (mandat):

1.4- Décrivez votre secteur et ses relations avec le reste de l'institution :

1.5- Responsabilités : (Vos principales responsabilités et celles de votre service)

1.6- Objectifs et problèmes professionnels :

1.7- Besoins en matière d'analyse et de données :

1.8- Critère de réussite du projet :

1.9- Nombre d'années d'expérience à ce poste :

1.10- Nombre d'années d'expérience dans le milieu universitaire :

ANNEXE F

«Questionnaire aux dirigeants»

Partie 2 : Utilisation des systèmes de gestion de l'UQTR

2.1-Voici les principaux systèmes de l'Université :

Veuillez cocher ceux que vous ou votre service utilisez à des fins:

- **d'extraction de données pour traitement ultérieur;**
- **de consultation (rapports avec sommation, moyenne, ...);**
- **de prise de décision.**

	Acronyme système	Brève description du système		Acronyme système	Brève description du système
☐	ADM	Admission en ligne	☐	API	Analyse et planification institutionnelle
☐	BOU	Gestion des bourses universitaires	☐	DAF	Dossier académique et financier de l'étudiant
☐	DCI	Direction de la coopération internationale	☐	DIP	Gestion des diplômations et de la collation des grades
☐	EVA	Évaluation de la qualité des enseignements	☐	FON	Système de la Fondationd e l'UQTR
☐	GAR	Gestion des activités et des ressources/td>	☐	GEN	Système général
☐	INF	Gestion des rubrique de nouvelles de l'en-tête et portail	☐	PCO	Portail de cours
☐	PER	Gestion des personnels	☐	RAD	Réquisitions de décision relative à la demande d'admission
☐	REC	Reconnaissance des acquis	☐	STG	Gestion des stages
☐	SUB	Demande de substitution	☐	TRI	Cours et programmes

2.2-Veuillez cocher les modules que vous utilisez le plus fréquemment :

Indiquez-en aussi la fréquence (jour/semaine/mois/session/annuelle):

- **d'extraction de données pour traitement ultérieur;**
- **de consultation (rapports avec sommation, moyenne, ...);**
- **de prise de décision.**

Système	Module	Tables		Fréquence	Période
Fondation (FON)	Campagne de financement	FON_CAMPAGNE	Consultation	☐	jour
			Extraction	☐	jour
			Décision	☐	jour
	Compte de fonds	FON_FONDS	Consultation	☐	jour
			Extraction	☐	jour
			Décision	☐	jour
	Promesse de don	FON_PROMESSE	Consultation	☐	jour
			Extraction	☐	jour
			Décision	☐	jour
	Sollicitation	FON_SOLICITATION	Consultation	☐	jour
			Extraction	☐	jour
			Décision	☐	jour
	Versement aux lauréats	FON_VERSEMENT	Consultation	☐	jour
			Extraction	☐	jour
			Décision	☐	jour

ANNEXE F

«Questionnaire aux dirigeants»

2.3- Est-ce que ces systèmes répondent bien à vos besoins d'informations

○ oui ○ non

Commentaires:

2.4- Utilisez-vous d'autre(s) système(s) non mentionné ci-haut

Acronyme système	Brève description du système	Type opération	Choix	fréquence	type
#1		Consultation	☐		jour
		Extraction	☐		jour
		Décision	☐		jour
#2		Consultation	☐		jour
		Extraction	☐		jour
		Décision	☐		jour

3.5-Est-ce qu'une personne intermédiaire s'occupe de préparer les données à des fins d'analyse ?

○ oui ○ non

Si oui, qui :

Partie 4 : Prise de décision

4.1-Citez 2 exemples concrects de décision budgétaire prises au cours de la dernière année.

Exemples #1:

Exemples #2:

4.2-Citez 2 autres exemples concrects de décision (autres que budgétaire) prises au cours de la dernière année.

Exemples #3:

Exemples #4:

ANNEXE F

«Questionnaire aux dirigeants»

Partie 3 : Utilisation d'autres sources

3.1-Avez-vous accès facilement aux données informatiques donc vous avez besoin pour votre prise de décision?

○ oui ○ non

Commentaires:

3.2-Idéalement, que souhaiteriez-vous avoir comme données informatiques pour décider plus facilement ? (avoir plus d'information, plus facilement, plus rapidement,...)

Commentaires:

3.3-Est-ce que présentement, vous effectuez une extraction de données afin de les traiter avec d'autres outils informatiques ? (Ex.: Excel, Word, Access,...)

○ oui ○ non

Si oui, lesquels :

Donnez des exemples:

3.4-Détenez-vous d'autres sources de données que celles provenant des syatèmes de gestion de l'UQTR ?

○ oui ○ non

Si oui, lesquels :

ANNEXE F

«Questionnaire aux dirigeants»

4.3-Précisez les caractéristiques des données recherchées.
(Ex.: moyenne cumulative, nombre de crédits réussits, ...)
Pour l'exemple #1 ci-haut:

Pour l'exemple #2 ci-haut:

Pour l'exemple #3 ci-haut:

Pour l'exemple #4 ci-haut:

4.4-Quelle est la période de temps couverte par les données recherchées.
(Ex.: 2 dernières années, dernière session, ...)

4.5-Contactez-vous un autre Service ou Département pour votre cueillette de données ?
○ oui ○ non

Si oui, précisez leur(s) nom(s) :

ANNEXE G

No.		CRITÈRES D'ÉVALUATION	Oracle WB 10g	JasperSoft	BO - Crystal Decisions	Informatica - PowerCenter	Cognos 8
		LES CARACTÉRISTIQUES FONCTIONNELLES					
1	CDC	**Préparation des données**					
		Soutien « clé en main » (valeur ajoutée) pour traitement CDC	Non	Non	Non	**Oui**	Non
2	ETL	**Nettoyage des données**					
		Analyse de la qualité de données (data profiling)	Oui / ?	Non	Non	Oui	Non
		Gestion de la qualité de données (par transformations ou règles)	Oui / ?	Non	Oui / ?	Oui / ?	Non
		Chargement des données					
		Analyse des impacts	Non	Non	**Oui**	**Oui**	Non
		Trace d'origine des données (data lineage)	Non	Non	**Oui**	**Oui**	Non
		Interface de gestion pour céduler les chargements ?	**Oui**	**Oui**	Oui	**Oui**	**Oui**
		Mécanisme d'avis d'erreurs de chargement ?	**Oui**	**Oui**	Oui	**Oui**	**Oui**
		Interface de suivi des opérations	**Oui**	**Oui**	?	**Oui**	**Oui**
		États des opérations (type tableau de bord)	Non	Non	Non	**Oui**	Non
		Transformation des données					
		Aggrégations, Filtres et Lookup, etc.	**Oui**	**Oui**	**Oui**	**Oui**	**Oui**
		Langage et interfaces de programmation (APIs)	**Oui**	**Oui**	**Oui**	?	**Oui**
			PL/SQL	Java ou Perl	?	?	Cognos Script et PL/SQL
		Gestion des modèles dimensionnels					
		La définition et la gestion des méta-données	**Oui**	Non	?	**Oui**	**Oui**
		Gestion des tables de fait et dimensions	**Oui**	**Oui**	**Oui**	**Oui**	**Oui**
		Gestion de la conformité des dimensions	Non	Non	Non	Non	Non
		Gestion des dimensions lentes (SCD)	**Oui**	**Oui**	**Oui**	**Oui**	**Oui**
		Gestion des faits en retard (late arriving facts)	Non	Non	Non	Non	**Oui**
		Intégration et opérations avec les bases Oracle	**Oui**	**Oui**	**Oui**	**Oui**	**Oui**
		Gestion de changement de clés primaires dans OLTP	Non	Non	Non	Non	Non
		Mise à jour de l'entrepôt					
		Représentation explicite de la dimension de temps	**Oui**	**Oui**	**Oui**	**Oui**	Non
		Support pour les modes ROLAP et MOLAP	**Oui**	Non	Non	Non	Non
		Facilité du déploiement de l'entrepôt ou des data marts	**Oui**	**Oui**	?	Non	**Oui**

No.		CRITÈRES D'ÉVALUATION	Oracle WB 10g	JasperSoft	BO - Crystal Decisions	Informatica - PowerCenter	Cognos 8
		LES CARACTERISTIQUES FONCTIONNELLES					
3	BI	**Les rapports et tableaux de bord**					
		Création de rapports adhoc	Non	**Oui**	Oui / ?	**Oui**	**Oui**
		Création de tableaux de bord	**Oui**	**Oui**	Oui / ?	Non	**Oui**
		Création de rapports paramétrisés	**Oui**	**Oui**	Oui / ?	Oui	**Oui**
		Objets graphiques intéressants (jauge, cartes, indicateurs, tableaux intelligents)	Non	**Oui**	**Oui**	Non	**Oui**
		Outils d'analyse					
		Moteur OLAP	Oui / ?	**Oui**	**Oui**	Non	**Oui**
		LES CARACTÉRISTIQUES NON FONCTIONNELLES					
4		**Environnement et interfaces**					
		Convivialité des interfaces utilisateurs ETL	**1**	**2**	**?**	**3**	**2**
		Convivialité des interfaces utilisateurs BI	**1**	**2**	**?**	**0**	**3**
		Support en utilisant des wizards	**?**	**1**	**?**	**1**	**2**
		Service « avant vente »					
		Qualité de la présentation	**0**	**3**	**1**	**3**	**3**
		Suivi de la part du représentant (à nos questions)	**0**	**3**	**0**	**3**	**3**
		ÉCHELLE DE COTATION :					
		Oui = la fonctionnalité est disponible					
		Non = la fonctionnalité n'est pas disponible					
		/? = Indéterminer (pas démontré lors de la démo)					
		0 = Inacceptable, 1 = Passable, 2= Bien, 3= Très bien					

COMPARAISON DES OUTILS - LOGICIELS LIBRES (PA10)

par Michel Charest

#	CRITÈRE D'ÉVALUATION	JasperSoft	Pentaho	Poids	Jasper Score	Jasper Total	Pentaho Score	Pentaho Total
k.	Convivialité des objets (constructs) et fonctionnalités en général	Moyen - Les opérations de base sont relativement facile à opérer, mais la majorité des fonctionnalités se font par un menu déroulant de genre "pop-up". Il serait mieux de prévoir une barre d'outil et des onglets pour faciliter l'utilisation car le menu "pop-up" devient désagréable.	Inconnu – pas testé ?	3	2	6		0
l.	Gestion des éléments (query et data items) de requêtes durant la création et gestion d'un rapport	Moyen - L'outil est limité lorsqu'on a des besoins de rapports plus poussés. Il devient donc nécessaire d'utiliser un logiciel séparé – iReports (qui doit être installé sur un poste de travail).	Inconnu – pas testé ?	2	2	4		0
m.	Ajustement et peaufinage des rapports actualisés	Moyen - Limité strictement sur le Web ... doit utiliser iReports (application à installer sur un poste de travail).	Inconnu – pas testé ?	2	2	4		0
n.	Rapidité et temps réponse des écrans	Bien - lorsqu'un rapport est créer (p.ex.: en glissant-collant des éléments) l'outil réagis relativement bien.	Inconnu – pas testé ?	3	3	9		0
			SOUS - TOTAL :			47		2
4)	ANALYSES ET EXPLORATIONS :	JasperAnalysis (utilise Mondrian)	Mondrian					
a.	Création de la structure principale du rapport d'analyse (tableau croisée)	Faible - gérer manuellement	Faible - gérer manuellement	2	1	2	1	2
b.	Utilisation des hiérarchies (niveaux et membres)	Moyen - convivialité laisse à désirer. On doit gérer des fichiers de configuration XML (il existe un outil séparer de genre "Wizard" pour créer ceux-ci, mais pas pas testé. (semble relativement limité). Ceci resoulève le même point du manque des gestion des métas-données (voir 2a). – De plus, pour l'instant, pas possible de partager des dimensions lorsqu'on configure les fichiers XML, dont pas de "stitch query"	Moyen - convivialité laisse à désirer. On doit gérer des fichiers de configuration XML (il existe un outil séparer de genre "Wizard" pour créer ceux-ci, mais pas pas testé. (semble relativement limité). Ceci resoulève le même point du manque des gestion des métas-données (voir 2a). – De plus, pour l'instant, pas possible de partager des dimensions lorsqu'on configure les fichiers XML, dont pas de "stitch query"	2	2	4	2	4
c.	Emplacement des niveaux ("Nesting", "Stacking")	Moyen - On peut faire du nesting (pas du stacking)	Moyen - On peut faire du nesting (pas du stacking)	2	2	4	2	4
d.	Triage	Bien - Un simple bouton	Bien - Un simple bouton	2	3	6	3	6
e.	Filtrage	Bien - Un simple bouton	Bien - Un simple bouton	2	3	6	3	6
f.	Ajouter de calculs (p.ex.: sous-totaux)	Inconnu ?	Inconnu ?	2		0		0
g.	Forage ("Drill", "Roll-up", "Slice", "Swap")	Moyen - Relativement facile à forer, mais interface n'est pas aussi convivial que les outils commerciaux.	Moyen - Relativement facile à forer, mais interface n'est pas aussi convivial que les outils commerciaux.	2	2	4	2	4
h.	Permet l'utilisation des cubes MOLAP ou outils d'analyse OLAP qui est "aggregate aware"	Moyen - cubes relationnelles seulement, mais support existe pour tables d'aggrégations	Moyen - cubes relationnelles seulement, mais support existe pour tables d'aggrégations	3	2	6	2	6
i.	Permet les requêtes MDX (Multi Dimensionnal Extension)	Bien - supporté par l'outil	Bien - supporté par l'outil	1	3	3	3	3

COMPARAISON DES OUTILS - LOGICIELS LIBRES (PA10)

par Michel Charest

#	CRITÈRE D'ÉVALUATION	JasperSoft	Pentaho	Poids	Jasper Score	Jasper Total	Pentaho Score	Pentaho Total
1)	INTÉGRATION DES DONNÉES (ETL) :	JasperETL	Pentaho Data Integration (Kettle)					
a.	Support "natif" pour Oracle CDC	Aucun	Aucun	2	1	2	1	2
b.	Utilisation de méta-données (p.ex.cadre dimensionnel)	Aucun	Aucun	2	1	2	1	2
c.	Acquisition de données (Extraction)	Bien - offre des constructs de connexions aux base de données et différents formats de fichiers (.csv, XML, etc.)	Bien - offre des constructs de connexions aux base de données et différents formats de fichiers (.csv, XML, etc.)	2	3	6	3	6
d.	Transformation des données (Transformation)	Moyen - Offre un ensemble d'opérateurs (composantes) qui sont liées en un flot pour effectuer des transformations sur les données. La librairie semble assez complète, mais n'est pas ne semble pas très intuitive à utiliser. De plus, l'usager doit se familiariser avec beaucoup de propriétés pour chacune des composantes.	Bien - Offre un ensemble d'opérateurs (composantes) qui sont liées en un flot pour effectuer des transformations sur les données. La librairie semble assez complète, également pas toujours intuitif à utiliser. Un avantage (par rapport à JasperETL) est que l'outil à des types de composantes (p.ex.: connexion, transforamtion et job) ce qui facilite la compréhension et l'utilisation des composantes.	3	2	6	3	9
e.	Chargement des données (Load)	Bien - Offre des méthodes pour charger des tables relationnelles (Oracle, SQLServer, etc.), ainsi que des fichiers de différents formats (.csv, XML, etc.)	Bien - Offre des méthodes pour charger des tables relationnelles (Oracle, SQLServer, etc.), ainsi que des fichiers de différents formats (.csv, XML, etc.)	2	3	6	3	6
f.	Création et gestion des flots (p.ex. parallélisme)	Moyen - offre un environnement graphique pour réaliser des flots ETL (liens et opérateurs provenant d'une boite d'outil/constructs). Typiquement, un flot comprends une "Job" qui est composé de un ou plusieurs opérateurs."	Moyen - offre un environnement graphique pour réaliser des flots ETL (liens et opérateurs provenant d'une boite d'outil/constructs). Typiquement, un flot comprends une "Job" qui est composé de un ou plusieurs opérateurs."	2	2	4	2	4
g.	Gestion des dimensions lentes (SCD) et des hierarchies déséquilibrées	Moyen - La documentation mentionne une composante (tOracleSCD) qui permet de gérer la gestion des dimensions lentes (SCD), mais elle n'a pas été testé. – Pas clair comment gérer les opérations DELETE ? Pas clair si les hierarchies déséquilibrées song gérées ?	Moyen - La documentation mentionne une composante (Dimension/Lookup Update) qui permet de gérer la gestion des dimensions lentes (SCD), mais elle n'a pas été testé. – Pas clair comment gérer les opérations DELETE. Pas clair si les hierarchies déséquilibrées song gérées ?	3	2	6	2	6
h.	Gestion des changements de structures	Faible - on devra intervenir en utilisant la gestion des erreurs haut-niveau dans l'ETL et en utilisant les fonctionnalités Oracle CDC.	Faible - on devra intervenir en utilisant la gestion des erreurs haut-niveau dans l'ETL et en utilisant les fonctionnalités Oracle CDC.	1	1	1	1	1
i.	Outils a base d'assistant ("wizard")	Aucun	Aucun	2	1	2	1	2
j.	Intégration avec gestion des méta-données (voir section 2)	Aucun	Aucun	1	1	1	1	1
k.	Permet l'exécution par céduler ou l'invite de commande (shell)	Moyen - il n'y a pas de céduleur, mais il est possible d'exécuter les jobs en utilisant l'invite de commande et un céduleur externe (tel que crontab en Unix).	Bien - comprends à la fois un céduleur intégré dans l'outil ETL pour exécuter les jobs et la possibilité d'exécuter des celles-ci par la ligne de commande (et un céduleur externe).	2	2	4	3	6
l.	Le langage script et librairie de fonctions	Moyen - On dépends fortement sur les options et fonctionnalités intrinsectes aux opérateurs. Malgré cela, il est possible de faire des appels externes (PL/SQL, etc.)	Moyen - On dépends fortement sur les options et fonctionnalités intrinsectes aux opérateurs. Malgré cela, il est possible de faire des appels externes (PL/SQL, etc.)	3	2	6	2	6
			SOUS - TOTAL :			46		51

COMPARAISON DES OUTILS - LOGICIELS LIBRES (PA10)

par Michel Charest

#	CRITÈRE D'ÉVALUATION	JasperSoft	Pentaho	Points	Jasper Score	Jasper Total	Pentaho Score	Pentaho Total
2)	GESTION DES MÉTA-DONNÉES :	JasperServer	BI Platform					
a.	Gestion des tables ("query subject" et table dérivée)	Aucun - gérer à très "bas-niveaux" en utilisant l'outil ETL seulement. À NOTER: Ceci est la grande faiblesse de ces outils. Ils n'offre pas un environnement intégré pour gérer les méta-données et la modèles (à publier aux usagers) de façon intégré et centralisé.	Aucun - gérer à très "bas-niveaux" en utilisant l'outil ETL seulement. À NOTER: Ceci est la grande faiblesse de ces outils. Ils n'offre pas un environnement intégré pour gérer les méta-données et la modèles (à publier aux usagers) de façon intégré et centralisé.	1	1	1	1	1
b.	Utilisation d'alias de tables	Aucun	Aucun	3	1	3	1	3
c.	Définition des hiérarchies (niveaux et membres)"	Aucun -	Aucun	3	1	3	1	3
d.	Gestion des boucles, "fan traps" et "chasm traps"	Aucun -	Aucun	2	1	2	1	2
e.	Gestion de l'intégrité des modèles (univers, package, etc.)	Aucun	Aucun	1	1	1	1	1
f.	Outils de support à la conception	Aucun	Aucun	2	1	2	1	2
g.	Gestion de la sécurité et le contrôle des accès	Aucun	Aucun	2	1	2	1	2
h.	Convivialité de l'interface	Aucun	Aucun	2	1	2	1	2
i.	Analyse des impacts et "Data Lineage"	Aucun	Aucun	2	1	2	1	2
			SOUS - TOTAL :			18		18
3)	GESTION DES RAPPORTS :	JasperReports and iReports	JFreeReports					
b.	Spécification des champs de table (query item)	Bien - sélectionner à partir d'une liste d'objets	Inconnu – pas testé ? À NOTER : À cause de difficultés rencontrés durant l'installation, il ne fut impossible de faire des essais avec cette composante.	1	3	3		0
c.	Spécification des champs calculés (variable ou data item)	Inconnu – pas testé ?	Inconnu – pas testé ?	1		0		0
d.	Spécification du triage de champ	Bien - Simple option dans un menu déroulant (pop-up menu)	Inconnu – pas testé ?	1	3	3		0
e.	Ajout d'un filtre	Bien - Simple option dans un menu déroulant (pop-up menu)	Inconnu – pas testé ?	1	3	3		0
f.	Création des sections	Inconnu – pas testé ?	Inconnu – pas testé ?	1		0		0
g.	Spécification d'un group	Bien - Simple option dans un menu déroulant (pop-up menu)	Inconnu – pas testé ?	1	3	3		0
h.	Actualisation des données	Bien - Un simple bouton	Inconnu – pas testé ?	2	3	6		0
i.	Visualisation et interactivité avec la structure du rapport	Moyen - N'offre pas un mode pour visualiser la structure d'un rapport (baseé sur des modèles prédéfinies seulement)	Inconnu – pas testé ?	2	2	4		0
j.	Possibilité de comparer divers versions d'un rapport	Aucun	Aucun	2	1	2	1	2

COMPARAISON DES OUTILS - LOGICIELS LIBRES (PA10)

par Michel Charest

#	CRITÈRE D'ÉVALUATION	JasperSoft	Pentaho	Poids	Jasper Score	Jasper Total	Pentaho Score	Pentaho Total
j.	Convivialité de l'interface	Moyen - fonctionnel, mais l'interface n'est pas aussi convivial à utiliser que les outils commerciaux. (p.ex. l'application de filtres).	Moyen - fonctionnel, mais l'interface n'est pas aussi convivial à utiliser que les outils commerciaux. (p.ex. l'application de filtres).	3	2	6	2	6
k.	Exploitation des données et forage en utilisant Microsoft Office	Aucun	Aucun	2	1	2	1	2
			SOUS - TOTAL :			43		43
5)	INDICATEURS DE PERFORMANCE :	JasperReports	JFreeReports					
a.	Gestion des évènements	Inconnu – pas testé ?	Inconnu – pas testé ?	2		0		0
b.	Création des tableaux de bord (Dashboard)	Aucun	Inconnu – pas testé ?	3	1	3		0
c.	Création des cartes de pointage (Scorecard)	Aucun	Inconnu – pas testé ?	2	1	2		0
			SOUS - TOTAL :			5		0
6)	DÉTAILS DE DÉPLOIEMENT :							
a.	Plateforme et système d'exploitation	Bien - AIX ou Windows (application basée fortement sur Java)	Bien - AIX ou Windows (application basée fortement sur Java)	2	3	6	3	6
b.	Permet d'utiliser des APIs de programmation (p.ex.: Java) pour faire de l'intégration sur mesure à nos systèmes.	Bien - librairies Java (et services Web ?)	Bien - librairies Java (et services Web ?)	2	3	6	3	6
			SOUS - TOTAL :			12		12
7)	SOUTIENT LOCAL et COMMUNAUTÉ :							
a.	Représentants sur place (Montréal et environs)*	Aucun - Compagnie Américaine (Californie)	Aucun - Compagnie Américaine (Floride)	2	1	2	1	2
b.	Accès à une communauté/base d'utilisateurs	Faible - Universités Américaine seulement.	Moyen - (Université de Montréal exploite ce produit)	2	1	2	2	4
			SOUS - TOTAL :			4		6
	Légende des scores : Inconnu=0, Faible=1, Moyen=2, Bien=3, Excellent=4		TOTAUX (sur maximum de 412) :	103		175		132
	Légende des poids (selon importance) : Moyen=1, Important=2, Très important=3		TOTAUX (pourcentage) :			42%		32%

ANNEXE I

«Devis pour l'achat d'un système d'intelligence d'affaires (BI)»

Par Michel Charest

Définition : Un système d'intelligence d'affaires (« Business Intelligence » en anglais) est un système d'exploitation des données d'une entreprise qui facilite la prise de décision. Ce système permet de regrouper les informations nécessaires à la constitution de rapports, d'analyses et de tableaux de bord, permettant ainsi aux dirigeants d'obtenir rapidement un portrait clair de la situation de l'entreprise. Les systèmes d'intelligence d'affaires permettent également d'intégrer dans un même système les informations provenant de plusieurs départements. Ainsi, ce genre d'intégration offre la possibilité de réaliser rapidement des rapports consolidés qui permettent une vue unifiée d'une entreprise. Voici les exigences de base que devra combler le futur système :

Exigences générales :

a) **Un portail d'intelligence d'affaire** - Le système devra permettre aux usagers d'accéder aux diverses fonctionnalités en utilisant un portail Web (p.ex. : gestion et création de rapports, d'analyse OLAP et de gestion de tableaux de bord).

b) **Le modèle d'accès aux usagers à base de licences** – Le système devra permettre jusqu'à 20 utilisateurs à accès concurrents (licences non nommées) et la possibilité d'ajouter jusqu'à 20 utilisateurs supplémentaires nommées (avec un accès privilégié, mais assignable de nouveau en tout temps à une autre personne).

c) **La plate-forme ou le système d'exploitation** – Le système devra fonctionner sur la plateforme UNIX (plus précisément soit AIX 5.x ou Linux Redhat Entreprise 4.x)

Exigences pour la composante ETL[6] :

d) **6 utilisateurs concurrents** – le système devra permettre jusqu'à six utilisateurs (de façon concurrente) d'utiliser la composante ETL afin d'effectuer des tâches de conception et d'exécution de « flots de traitement » qui permettront le chargement efficace d'un entrepôt de données (compatible avec une base de données Oracle 10g et plus).

e) **Intégration avec système CDC[7] Oracle** – La composante devra permettre une intégration facile avec le produit CDC Oracle. Plus précisément, celle-ci doit permettre l'alimentation efficace de l'entrepôt, par l'entremise du produit CDC Oracle, pour toutes les opérations transactionnelles (p.ex. : les ajouts, les suppressions et mise à jour des données sur le côté opérationnel).

f) **Support pour les procédures stockées Oracle PL/SQL[8]** – L'environnement ETL devra permettre l'exécution de nos procédures stockées PL/SQL.

---

[6] ETL = Extract, Transform and Load (en anglais)
[7] CDC = Changed Data Capture (en anglais)
[8] PL/SQL = Procedural Language / Structured Query Language (en anglais)

ANNEXE I

«Devis pour l'achat d'un système d'intelligence d'affaires (BI)»

g) **Convivialité des interfaces** – Cette composante devra permettre à l'usager de facilement définir des « flots de traitement » (p.ex. : ajout des connexions source et cible, ajout des opérateurs de transformations pour filtrer et nettoyer les données, gestion des dimensions lentes, gestion des clés de source opérationnelle et des clés de l'entrepôt de données, séquenceur, exécuter de dépanner les flots de traitements, etc.)

Exigences pour la composante de gestion des modèles et métadonnées :

h) **Analyse des impacts** – Le système devra offrir des outils qui permettront aux concepteurs de gérer les impacts potentiellement problématiques (p.ex. : tels que l'invalidation des rapports préalablement conçus suite aux modifications qui seront apportées à la structure de l'entrepôt de données).

i) **Analyse de provenance des données (« data lineage »)** – Le système devra offrir des outils qui permettront aux concepteurs d'analyser la provenance des données de l'entrepôt de données (p.ex. : sources sur le côté opérationnel).

j) **3 utilisateurs** – devra permettre jusqu'à trois utilisateurs (de façon concurrente) d'utiliser la composante de modélisation de la structure de l'entrepôt de données (c'est-à-dire les modèles dimensionnels).

k) **Résolution des problèmes de jointures** – Le système devra offrir des outils qui permettront (aux concepteurs de l'entrepôt de données) de gérer des problèmes de jointures de tables reconnues qui peuvent surgirent lors de l'exploitation d'un entrepôt de données (p.ex. : boucles fermés, « fan-outs » et « chasm traps »).

Exigences pour la composante de création de rapports :

l) **La comparaison de différents scénarios ou rapports** – le système devra offrir une interface sur le web qui permet à l'usager de facilement réaliser et d'examiner divers scénarios de rapports « ad hoc ». Par exemple, l'interface web permettra à l'usager de consulter rapidement chacun des tableaux ou rapport en utilisant des « onglets ». De plus, il devrait être possible de sauvegarder l'ensemble des scénarios sous la forme d'une session d'exploration ou d'un dossier afin de faciliter leur consultation à un moment antérieur.

m) **La « forabilité » intégrée dans les rapports** - il devra être intuitif et facile pour l'usager de « forer » au sens OLAP[9], au besoin, lorsque celle-ci visualisera un rapport d'affaires (ayant été réalisé en utilisant des données hiérarchiques). Ce besoin est noté, car certains systèmes d'intelligence d'affaires n'offrent pas une vue intégrée des rapports « ordinaires » et des rapports du genre « OLAP » (c'est-à-dire, l'usager doit basculer vers un mode ou l'autre et n'a pas accès aux deux types de rapports ou fonctionnalités à la fois).

[9] OLAP = On-Line Analytical Processing (en anglais)

ANNEXE I

«Devis pour l'achat d'un système d'intelligence d'affaires (BI)»

n) **Convivialité des interfaces** – Le système doit permettre aux usagers de facilement réaliser des rapports en utilisant un navigateur web qui leur donnera une expérience de travail (p.ex. : le « look and feel », le temps réponse des opérations, la familiarité des boutons et options de menu, etc.) » qui se rapproche aux applications que l'on retrouve fréquemment sur un poste de travail standard (p.ex. : Microsoft Word, Excel, PowerPoint, Visio, etc.).Exigences pour la composante d'analyses OLAP :

o) **Les requêtes MDX**[10] - Le moteur OLAP doit permettre aux concepteurs de modèles dimensionnels d'exploiter les requêtes MDX soit lors de la création de rapports ou durant le dépannage du système.

p) **L'analyse ROLAP**[11] - Le système devra permettre aux concepteurs de l'entrepôt de réaliser des cubes de type relationnel (modèles dimensionnels), et non seulement des cubes propriétaires (MOLAP[12]).

q) **Convivialité des interfaces** – voir les points (m) et (n) ci-dessus.

Exigences pour la composante de création des tableaux de bord :

r) **Permet les analyses « What-If »** - il devra être possible de réaliser un tableaux de bord contenant des jauges (variables) qui permettent à l'usager de simuler différents scénarios (changer les valeurs des jauges). Ceci aura pour effet de permettre à l'usager de visualiser les impacts (négatif ou positifs) sur divers indicateurs contenus dans le tableau de bord.

Fournisseurs recommandés :

- Business Objects
- Cognos
- SAS

Estimation des coûts :

- Achat de base répondant aux exigences - approximativement 80 000$
- Frais d'entretien (20% par année)
- Frais pour les licences nommées supplémentaires (accès à toutes les fonctionnalités) – approximativement 750$ par licence

[10] MDX = Multi-Dimensional Extension (an anglais)
[11] ROLAP = Relational On-Line Analytical Processing (en anglais)
[12] MOLAP = Multi-Dimensional On-Line Analytical Processing (en anglais)